RESEARCHERID
THOMSON REUTERS
[INIT: January 2017]
ISSN 2348 – 8034
Impact Factor- 4.022
GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

APPROACHES IN PREDICTING URBAN AIR QUALITY IN BANGALORE CITY AND COMPARATIVE ANALYSIS OF PREDICTIVE MODELS USING DATA MINING TOOL

Asha N[*1] & Dr M P Indira Gandhi[2]
[*1]Research Scholar, Mother Teresa Women's University, Kodaikanal
[2]Department of Computer Science, Mother Teresa Women's University,Kodaikanal

ABSTRACT

The massive increase in the emission of air pollutants by different monitoring stations of a metropolitan city on a daily basis have to be monitored and prediction of air quality is essential to protect human health and environment. Data collected in large scale by monitoring stations is rich in data but with poor information, also fast growing tremendous amount of data is collected and stored in large databases. So the data have to be mined with useful information for decision making policies. Therefore the hidden patterns have to be extracted and proper data analysis is required. Data mining is concerned with finding hidden patterns on large scale data. The Predictive techniques of Data mining analyzes huge datasets to extract models describing important data classes in order to predict future data trends. This paper focuses on Regression and Neural Network Techniques that works on large disk resident data to predict air quality index using Weka as a data mining tool.

Keywords- *Air quality index, Predictive Techniques, Regression Techniques, Neural Network Techniques, Weka,*

I. INTRODUCTION

Data mining also known as Knowledge discovery in data mining (KDD) assists in transforming implicitly stored large databases into useful information and knowledge. The KDD techniques are required for digging useful knowledge from the repository of air pollutants that are collected at different monitoring stations of a metropolitan city. The real time monitoring system are installed in 15 places of Bangalore city that monitors and displays air pollution parameters such as Sulphur Dioxide (SO2), Oxides of Nitrogen (NO2), and Respirable Suspended Particulate Matter (RSPM / PM10) which indeed have polluted the city with increasing rate of Vehicles and Industries. The five years annual average air quality index of each pollutant of several stations is collected. This continuous emission of air pollutant data has led to the extraction of useful information by applying Predictive Data Mining Techniques. The air pollutant dataset are identified with continuous values SO2, NO2, and (RSPM / PM10) that are emitted regularly on a daily basis at different monitoring stations. These Continuous values can be predicted using Linear Regression techniques , Feed Forward Neural networks models to predict air quality index of SO2, NO2 and PM10.Linear Regression is one of the powerful and elegant method for estimating the values of continuous target variable. The model developed by Linear Regression technique is a straight line that is used to approximate the relationship between a single continuous predictor variable and a single continuous response variable. Multilayer Feed Forward Networks is the type of Neural Network on which the Back Propagation algorithm performs by describing the network topology for a given dataset. Several data mining tools are used such as Weka, R programming, and Rapid Miner, etc, for large data sets.

II. CASE STUDY

The Bangalore capital of Karnataka state known as garden city is prone to air pollution due to uncontrolled growth of vehicular population and wrong sitting of industries. The KSPCB (Karnataka state pollution control board) is monitoring ambient air quality (AAQ) of Bangalore city at 15 locations using manual equipments under National Ambient Air Quality Monitoring Programme (NAMP) covering Industrial Area, Mixed Urban Area and Sensitive Area. The monitoring stations store air pollutants on a daily basis. The average annual archived data is collected from KSPCB website which is composed of PM10, SO2 and NO2 and air quality index of each pollutant from 2011-2015 are used for analysis and prediction. The air quality index of each pollutant is a numerical which determines

the health impacts and breathing discomforts in an polluted environment. It is classified as Good if it is between 0-50 (minimal impact), Satisfactory ranges 51-100(Minor breathing discomfort to sensitive people), Moderate ranges 101-200 (Breathing discomfort to the people with lungs), Poor ranges 201-300(Breathing discomfort to people on prolonged exposure).

Weka(Waikato for environmental knowledge analysis) is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. The data obtained from KPSCB are stored in Microsoft Excel sheet with CSV file format. The data contains 75 instances and 6 attributes. The pollutants and air quality index are taken as data fields. The data is preprocessed and stored in weka explorer in arff file format.

III. LINEAR REGRESSION TECHNIQUE

Linear Regression Technique is modeled using a straight line with a random variable Y called the response variable and another random variable X called the Predictor variable. Therefore the straight line is

$$Y = \alpha + \beta X \quad \text{eq-1}$$

Here, the variance Y is assumed to be constant, α and β are regression coefficients specifying the Y intercept of X. These coefficients can be solved by the method of least squares which minimizes the error between the actual data and the estimate of the line , since linear regression works on two parameters the sample data of SO2 and its air quality index, NO2 and its air quality index, PM10 and its corresponding air quality index are represented as X and Y in the form (x1, y1), (x2, y2),……(x3,y3), individually ,then the regression coefficients can be estimated using least square method. The coefficients α and β often provide good approximation to predict air quality index Y for the predictor X. The Linear Regression model is applied to individual pollutants PM10, SO2 and NO2 and its corresponding air quality index, to extract a model that fits the straight line. The following are the observations made for each of the pollutant with the training set.

Table 1: Evaluation Summary of the training set

Pollutant	**Correlation Coefficient**	**Mean Absolute Error**	**Root Mean Square**	**Relative Absolute Error**	**Root Relative Squared Error**
PM10,PM10 index	0.9912	3.7599	5.0181	5.0181%	13.2318%
SO2,SO2 index	0.9834	1.2854	1.9131	21.4485 %	18.1403 %
NO2,NO2 index	0.8566	1.8988	6.7391	23.3604 %	51.5961 %

The linear regression model developed for each pollutant is tested with the test input. The test set is given as an input to the model and the predicted results are plotted in a graph with observed air quality index verses predicted air quality index of SO2, NO2 and PM10.

ISSN 2348 – 8034
Impact Factor- 4.022

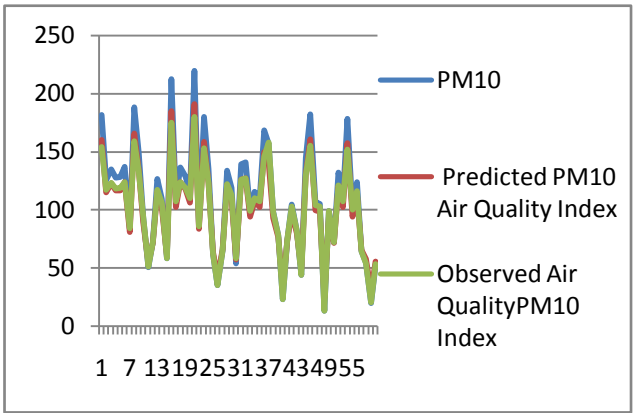

Figure 1: Test data set line plots Observed and Predicted Air Quality Index values of PM10

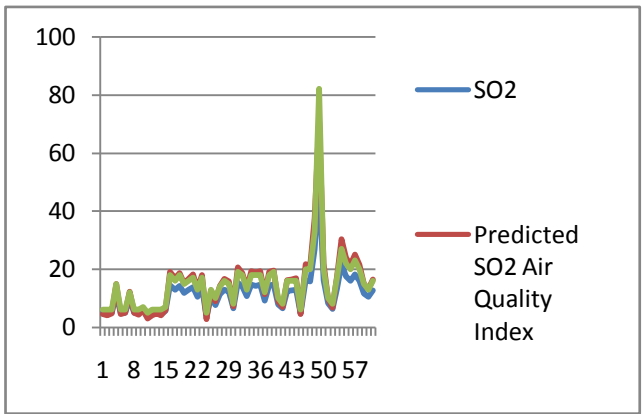

Figure 2: Test data set line plots Observed and Predicted Air Quality Index values of SO2

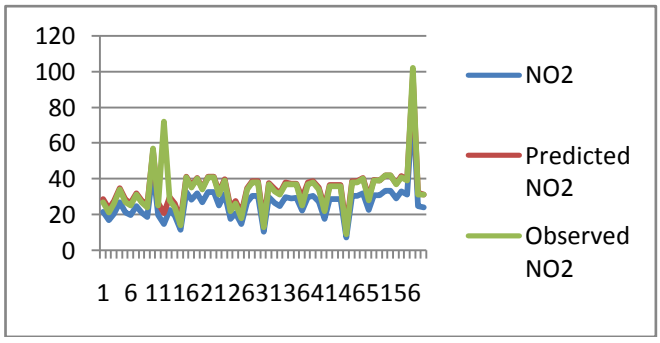

Figure 3: Test data set line plots Observed and Predicted Air Quality Index values of NO2

IV. MULTIPLE LINEAR REGRESSION MODELS

Multiple regressions are an extension of linear regression that involves more than one predictor variable. It allows response variable Y to be modeled as a linear function of a multidimensional feature vector. Multiple regression model based on two or more predictor variables X1, X2 is computed as follows.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

eq-2

ISSN 2348 – 8034
Impact Factor- 4.022

Here, the variance Y is assumed to be constant and β are regression coefficients specifying the y intercept of X1,X2 and so on. The Multiple Linear Regression model is applied to all pollutants PM10, SO2 and NO2 and its corresponding air quality index, to extract a model that fits the straight line for multiple attributes. The following are the observations made for all the pollutants. The Air quality index is a response variable Y dependent on the air pollutants and its air quality index which results with the following model and the table shows the observation made for the training set.

Air quality index of all pollutants = 1.0477 * PM10 + -0.2735 * PM10 Index + 0.3557 * So2 + -0.1049 * So2 Index +-0.138 * No2 +

0.1429 * No2 index +-8.2299

The Multiple Linear regression model is then tested with the test set to predict the Observed and Predicted Air Quality Index of all the Pollutants as shown in the following graph.

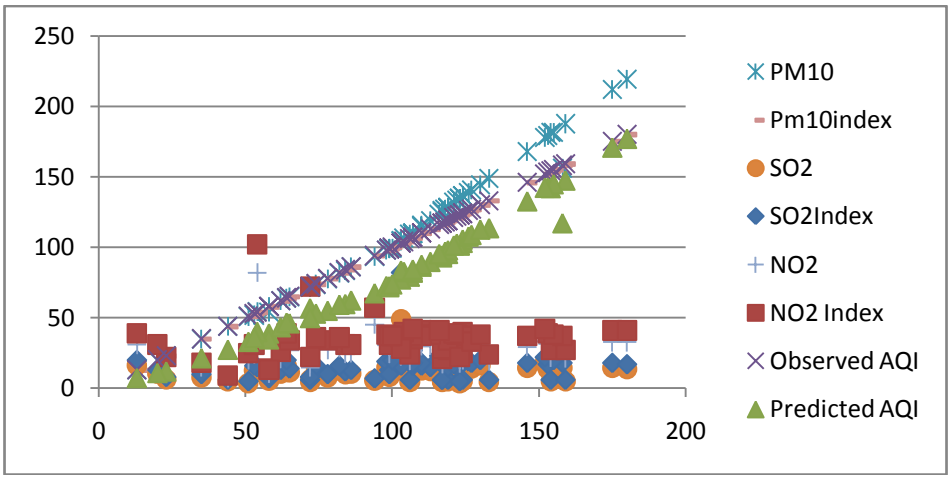

Figure 4: Test data set scatter plots Observed and Predicted Air Quality Index of all Pollutants

V. MULTILAYER FEED FORWARD NEURAL NETWORKS

It is a network of three layers such as input layer, hidden layer and output layer. The input corresponds to the attributes measured for each training sample. The inputs are fed simultaneously into layer of units making up the input layer. The weighted outputs of these units are fed to the second layer known as the hidden layer. The weighted outputs are inputs to the output layer which emits network predictions for given sample[6].The input layer consists of inputs such as PM10,PM10 index,SO2,SO2 index and NO2 , NO2 index to predict the air quality index in the output layer of Bangalore city. The network is trained with different networks to get the acceptable output.

VI. BACK PROPAGATION ALGORITHM

Back Propagation learns iteratively with a set of training sample by comparing the network predictions for each sample with actual known class label, the weights are modified such that the mean square error is minimized between network predictions and actual class, these modifications are made in the backward direction form output layer to the hidden layer .The model developed by Multilayer Feed Forward network contains Air quality index of all pollutants. The weights in the network are initialized to small random numbers and propagated to the hidden layer. The net input to the hidden layer and output layer is computed as a linear combination of its input. Given the unit j in hidden or output layer, the net input I_j to the j is computed as

$$Ij=\sum_i W_{ij}O_i+\Theta_j \quad \text{eq-3}$$

Where, Wij is the weight of the connection form unit naïve in the previous layer to the j.Oi is the output of unit I from the previous layer and Θj is the bias of the unit. Each unit in the hidden layer and output layer takes its net input and then applies an activation function called the sigmoid function. The error is propagated backwards by updating the weights and biases to reflect the error of the network predictions. Thus the network is trained so as to get the stable performance error. The best MLP Network model was the optimum found by the iterative process [1]. The trained network was tested with test data set so as to get the observed and predicted values of the air quality index. The following were the observations made by applying Back Propagation algorithm to the network.

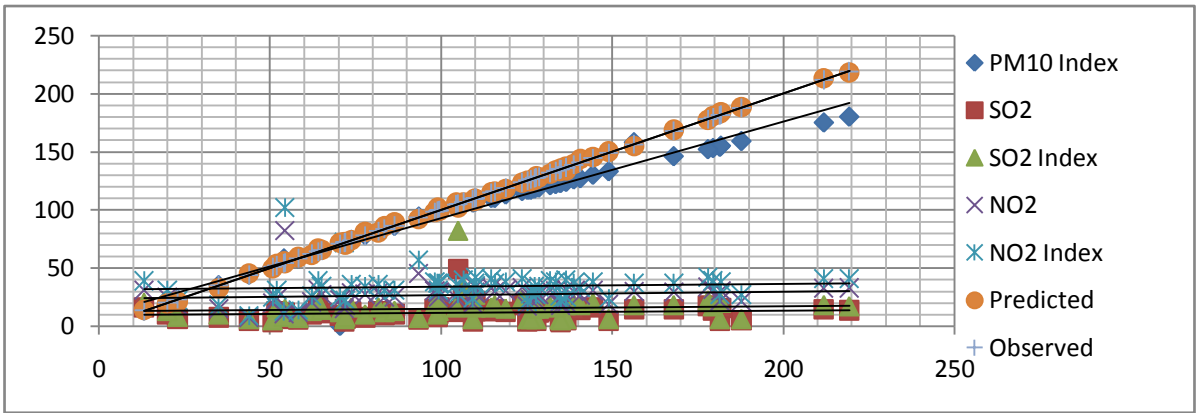

Figure 5: Test data set scatter plots Observed and Predicted Air Quality Index of all Pollutants By Back Propagation Algorithm

VII. COMPARATIVE STUDY OF THE DEVELOPED MODELS:

The Linear Regression technique was applied on two attributes such as PM10, SO2 and NO2 and corresponding index values so as to get the air quality index of that particular pollutant to know the health hazards caused by each pollutant in Bangalore city. The Multiple Linear Regression was applied with all Pollutants and its Air quality index so as to known the overall pollutant affecting the Bangalore city. The Back Propagation Algorithm in turn was also targeted with the predictions of Air quality index of the Bangalore city by considering all the pollutants and its index. The above Models were been used to Predict the air quality index that is seriously affecting the human health and environment. The statistical observations of all the developed models are compared with their level of predictions in terms of accuracy and relative error.

Table 2: Comparative study of Multiple Regression and MLP

Models	Correlation Coefficient	Mean Absolute Error	Root Mean Square	Relative Absolute Error	Root Relative Squared Error
Multiple Linear Regression	0.6559	31.4361	39.9718	69.0908 %	75.4877 %
Multilayer Perceptron Model	0.9995	1.2171	1.4764	3.3019	3.1945

From the above statistical observations, it could be identified that the relative error in Multiple Linear regression is more compared to MLP(Multilayer Perceptron) while making predictions, the MLP technique results with more accurate and good Predictions compared to Multiple Linear regression. The Multiple regression predictions are relatively less since the problem for fitting the co-efficient are solved using least squares. The following graph depicts the relative error, Observed AQI and Predicted AQI of both the models.

ISSN 2348 – 8034
Impact Factor- 4.022

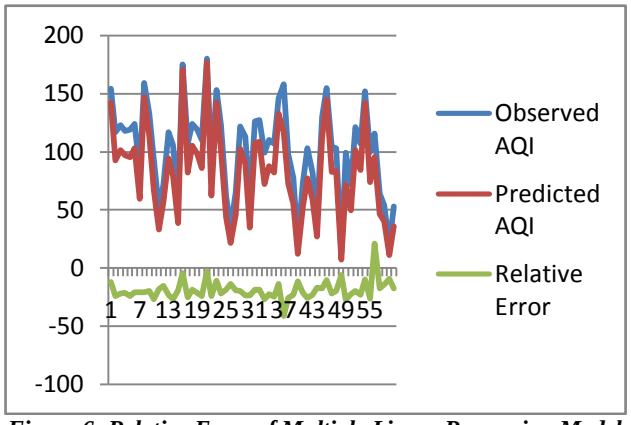

Figure 6: Relative Error of Multiple Linear Regression Model

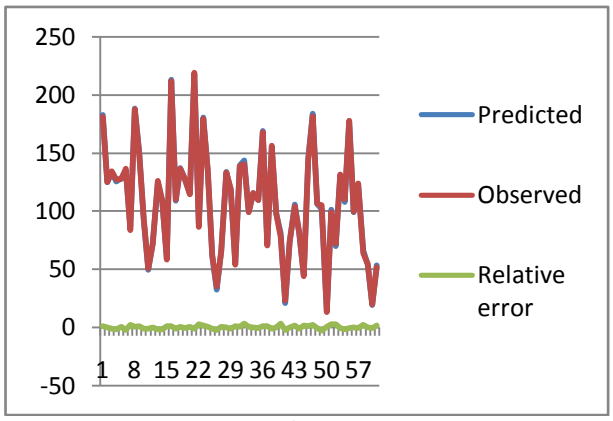

Figure 7: Relative Error of Multilayer Perceptron Model

From the above graph, it could be predicted that in MLP Predicted Air quality index values is almost near to the Observed air quality index values and hence the relative error is less compared to Multiple regression Model as observed.

VIII. CONCLUSION

The Data mining techniques, used to predict the air quality index for any given pollutant by constructing a model. Also it is observed that maximum pollutant that is affecting the Bangalore city is Pm10 with a maximum air quality index which is a major pollutant released by vehicles. The air quality index standards are classified as Moderate, Satisfactory, Good and Poor quality of air that is not permissible for a good breathing environment .The prediction provides decision making capability to the government to take proper action against the alarming rate of increase in vehicular population that has led to serious increase in PM10 concentration. This work paves way for further analysis of predictions made by MLP that can be classified into Moderate, Satisfactory, Good and Poor which can be given as an input to Decision tree and Naïve Bayes Algorithm to process nominal data and identify the highly polluted place in Bangalore city.

REFERENCES

1. *S. Christy, Dr. V. Khanaa "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 February 2016.*
2. *Jeremy E. Diem, Andrew C. Comrie "Predictive mapping of air pollution involving sparse spatial Observations " Department of Anthropology and Geography, Georgia State University, Atlanta The University of Arizona, Tucson, AZ 85721, USA Received 3 July 2001; accepted 16 October 2001.*
3. *Krzysztof Siwek, Stanislaw Osowski "Data mining methods for prediction of air pollution -Extended summary "Warsaw University of Technology, POLAND.*
4. *kspcb.kar.nic.in / The Karnataka State Pollution Control Board for Prevention and Control of Water Pollution constituted by the Government of Karnataka.*
5. *Article "Two-wheelers are biggest pollutants in Bangalore" Saswati mukherjee b| tnn | apr 14, 2013, 04.09 am ist.*
6. *"Data Mining concepts and techniques" by Jiwaei han and Micheline Kamber.*
7. *"Data mining methods and Models" by Daneil T Larose.*