Prediction and Analysis of Pollution Levels in Delhi Using Multilayer Perceptron

Aly Akhtar, Sarfaraz Masood, Chaitanya Gupta and Adil Masood

Abstract Air Pollution is a major problem faced by humans worldwide and is placed in the top ten health risks. Particulate Matter (PM10) is one of the major parameters to measure the air quality of an area. These are the particulate matter of the size 10 μm or less suspended in the air. PM10 occur naturally from volcanoes, forest fire, dust storms etc., as well as from human activities like coal combustion, burning of fossil fuels etc. The PM10 value is predicted by multilayer perceptron algorithm, which is an artificial neural network, Naive Bayes algorithm and Support Vector Machine algorithm. Total of 9 meteorological factors are considered in constructing the prediction model like Temperature, Wind Speed, Wind Direction, Humidity etc. We have then constructed an analysis model to find the correlation between the different meteorological factors and the PM10 value. Results are then compared for different algorithms, which show MLP as the best.

Keywords Artificial neural network · Naive Bayes · Support vector machine · Multilayer perceptron (MLP) · Correlation · PM 10 · Air quality index · Delhi

A. Akhtar · S. Masood (✉) · C. Gupta
Department of Computer Engineering, Jamia Millia Islamia, New Delhi, India
e-mail: smasood@jmi.ac.in

A. Akhtar
e-mail: alyakhtar94@gmail.com

C. Gupta
e-mail: cgupta319@gmail.com

A. Masood
Department of Civil Engineering, Al-Falah University, Faridabad, India
e-mail: adil.engg.cvl@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

S.C. Satapathy et al. (eds.), *Data Engineering and Intelligent Computing*,
Advances in Intelligent Systems and Computing 542,
DOI 10.1007/978-981-10-3223-3_54

1 Introduction

The capital of India—Delhi, according to a survey by World Health Organization (WHO) in 1600 cities around the world, has the worst air quality among the major cities [1]. About 1.5 million people are estimated to die every year because of air pollution and it is the fifth largest killer in India. India tops the list of death rate from chronic diseases and asthma in world. In Delhi, poor air quality is damaging the lungs of 2.2 million or 50% of all children. The major sources of pollution include Primary Pollutants e.g. CO, SO_x, PM etc., Secondary Pollutants e.g. H_2SO_4, O_3 etc., Vehicular Emissions and Industrial Pollution.

Among all the pollutants, Particulate Matter is a significant pollutant due to its effect on people's health when found in higher concentration than normal. The Study of the relationship of various meteorological parameters and the corresponding air pollution levels may not be very encouraging, as it does not affirm any interrelation between the parameters. However, the air pollutant levels usually respond to all metrological variables representing an air mass. This study attempts to examine the relation between the various meteorological features and urban air pollutant (Particulate Matter), and understand the relation between the local meteorology, atmospheric circulation and the concentrations of tropospheric air pollutants.

2 Literature Review

Air Quality Forecasting is a hot topic nowadays and we have more and more people working on it every coming day so as to come up with results that would in turn facilitate the government to have a greater control over the levels of Air Pollution. Some of the recent works in this field done globally have been discussed in this section.

2.1 Related Work

In [2], there is a study on the relationship between air pollution and individual meteorological parameters for predicting the Air Quality Index. However, they do not explain the interrelation between the parameters as the level of the air pollutants usually respond to all of meteorological variables representing an air mass. Studying individual meteorological factors is not sufficient as there are numerous meteorological variables that can have an effect on air quality.

The work done in [3] considers the cumulative effects of meteorological factors on PM 2.5. However this work includes only a few of the factors and misses out factors like wind speed and wind direction which are pretty important

meteorological factors and should be considered while investigating the trends with PM 10 values.

The work in [4] extensively studies the cumulative effects of a range of meteorological factors on the value of PM 1.0 and ultrafine particles (UFP). In this work PM 2.5 or PM 10 were not taken into account. UFP are particulate matter of nano scale size (less than 100 nm in diameter). As these particles are far smaller than PM 2.5 and PM 10 hence regulations do not exist for this size class of ambient air pollution particles. Due to this, UFP are not considered to be true indicatives of the Air Quality index. For a better understanding of an area's pollution, it is necessary to study either PM 2.5 or PM 10 along with UFP and PM 1.0.

In [5], the effects of meteorological conditions have been studied on the content of Ozone in the atmosphere. Global greenhouse effects are a proof that Carbon dioxide is one of the most harmful gases in the world.

The above discussed factors indicate that there exist some loopholes in the existing studies which account for incomplete or partial factors, so it is important to consider as much input factors as possible so as to increase the universal acceptability of the designed model and to achieve accurate results. The goal of this work is to design a neural network model that would accurately predict the level of PM 10 for Delhi and consequently analyze the pattern of PM 10 with variation of the input factors by studying the correlations between each individual factor and a combination of input factor and PM 10.

3 Design of Experiment

The data set collected for this work, consists of meteorological data for 396 days for the Delhi city. The idea is to use various machine learning techniques such as Multilayer Perceptron, Support Vector Machines and Naïve Bayes Classifier for the purpose of model construction and analysis.

3.1 Dataset

The Meteorological data was collected from Control Pollution Central Board (CPCB) [6]. The CPCB is the division of government of India under the ministry of Environment and Forests which tracks and researches on air quality in India. CPCB does the monitoring of meteorological parameters along with the monitoring of air quality at various centers in Delhi state as shown in the Fig. 1. For this work the R.K. Puram center was selected, as it is a densely populated area with many residential complexes, schools and various government buildings in its vicinity.

The data consisted of values for 396 days starting from 1st May 2015–1st June 2016 which had a total of 9 parameters: Average Temperature, Minimum Temperature, Maximum Temperature, Humidity, Pressure, Max Wind Speed, Visibility

Fig. 1 Map of Delhi [1] with marked location of data collection center

and Wind Speed, wind direction and PM 10. The data was fragmented into two parts: training dataset and the test dataset. Training data set has 316 sample observations and the test data has 80 sample observations. Each point represents the meteorological condition of a specific day in Delhi City at the selected center. A brief description of some of the parameters used in this work is as follows:

Average Temperature: Temperature affects the air quality due to the temperate inversion: the warm air above cooler air acts as a lid, thereby trapping the cooler air at the surface and suppressing vertical mixing. As pollutants from the vehicles, the industries, and the fireplaces are emitted into the air, this inversion phenomenon entraps the pollutants along the ground.

Average Wind Speed: Wind speed is a major player in diluting pollutants. Generally, scattering of the pollutants takes place due to strong winds, whereas light winds generally cause stagnant conditions, making pollutants to build up over an area.

Average Relative Humidity: Relative humidity (RH) is the ratio of the partial water vapor pressure to the equilibrium water vapor pressure at the same temperature.

Atmospheric Pressure: Atmospheric Pressure has a positive correlation with value of PM 10, this means that if the atmospheric pressure is high in an area then it is bound to be more polluted than an area having less atmospheric pressure.

Wind Direction: Wind Direction plays a major role in dispersion of the pollutants from one place to another. The major pollutants that enter Delhi are from Rajasthan i.e. South-West Direction.

Average Visibility: In meteorology, visibility is a measure of the distance at which a light or an object can be clearly distinguished. Average visibility is inversely proportional to pollution, this means that if visibility is high then it is very likely that value of PM 10 is low and vice versa.

3.2 Prediction Models

We used 3 Machine Learning techniques i.e. the Multilayer Perceptron (MLP), the Support Vector Machine (SVM) and the Naive Bayes method, for the purpose of building a predictor. The PM10 level were classified into "High" (>100 μg/m^3) and "low" (≤ 100 μg/m^3) categories on the basis of Air Quality Level standard in Delhi [7], which is 100 μg/m^3, and is considered to be a mild level pollution.

Multilayer Perceptron: It is a feed forward artificial neural network model that takes in a set of input data and maps it onto a set of outputs. MLP utilizes back-propagation for training a network, which is a supervised learning technique. In this work, the MLP consists of four layers, with 2 hidden layers (Fig. 2).

Support Vector Machines: SVM's are models of supervised learning technique, which can be used for data analysis using the classification or regression

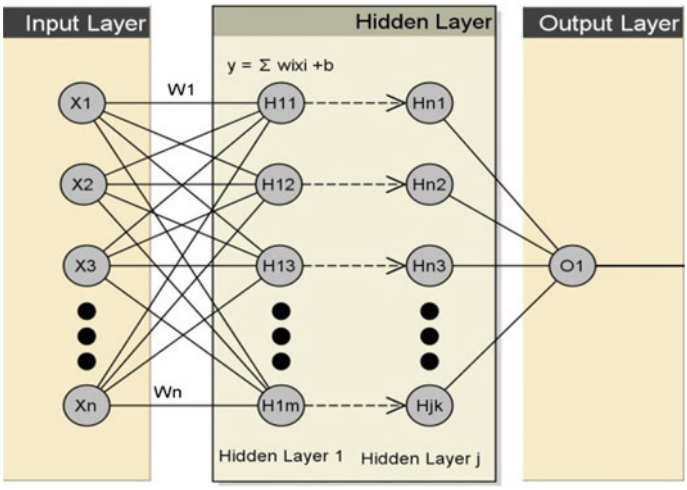

Fig. 2 Multilayer perceptron architecture

analysis. The model of SVM created is such that, all the data points are considered points in a space and these points are divided into clear categories as wide as possible.

Naïve Bayes: In machine learning, naive Bayes classifiers are a group of simple probabilistic based classifiers, which use the concept of Bayes Theorem.

Naïve Bayes is a set of algorithms, rather than a single algorithm whose aim is to consider each feature of the data set as an individual and independent entity, rather that every feature being connected or sharing it with each other.

Correlation. It is the dependence of any two or more variable on each other in a statistical relationship [8]. In this work the Pearson Coefficient has been used to find the correlation between each input factor and PM 10 separately. The Pearson correlation coefficient is the linear correlation whose value +1 refers to a total positive correlation, 0 refers to no correlation, and −1 refers to a total negative correlation.

3.3 Tools Used

For this work, the Python programming language was selected along with the following libraries:

- Pandas for performing data processing tasks such as data cleaning and normalization.
- Theano served as the base library for implementation of multilayer perceptron.
- Sklearn was used for implementing Naïve Bayes and SVM algorithm.
- Keras library that runs on top of theano, which helps us execute different algorithms.
- Scipy was used for determining correlation among the variables.
- Matplotlib for plotting graphs.

4 Analysis of Data and Correlation

It is evident from Fig. 3 that PM 10 and Wind direction show a positive correlation. The wind speed in association with the wind direction plays a major role in the dispersion of the pollutants. A dip in the concentration of the PM 10 levels in Delhi can be observed near the 60th day mark. Here the wind direction was mostly from the south or southeast direction and the wind speed was high, the pollutants dispersed quickly. Whereas around 217th day of the observation, where the PM10 value is at peak, it can be observed that the wind came from the south-west direction and the wind speed was quite slow, easing for the pollutants to settle around the surface and increase PM 10 levels.

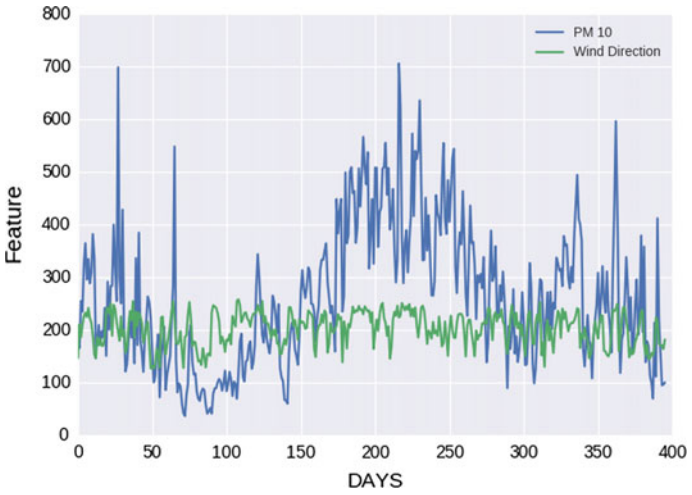

Fig. 3 Maximum positive correlation observed between PM10 and wind direction

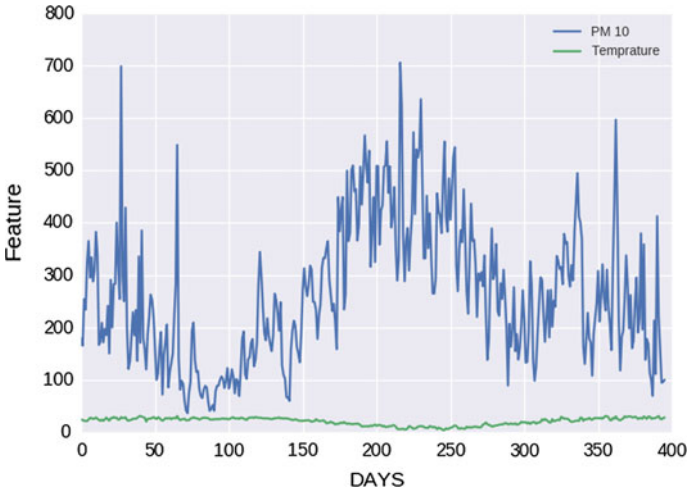

Fig. 4 Maximum negative correlation observed between PM10 and temperature

In Fig. 4, the clear effect of negative correlation between PM 10 and Temperature can be observed. As the temperature increases, the value of PM 10 will decrease and vice versa. Hence we can say that during summers the value of PM10 is less than what we observe during the winters, where the temperature is low.

5 Results

Classification-based predictions for test samples can be measured in many ways. One of the most straightforward of these measures is accuracy, which refers to the percentage of the samples that are correctly predicted. But, it may not be sufficient, hence precision, recall and f-measure value were also evaluated in this work.

5.1 Prediction

The prediction performance for the different models can be summarized as (Table 1):

Table 1 Results obtained for the various experiments

Technique	Accuracy (%)	Precision	Recall	F-measure
MLP	98.1	0.98	0.95	0.97
SVM	92.5	0.92	0.90	0.91
Naive Bayes	91.25	0.90	0.87	0.89

Table 2 Data of 20 sample days for wind direction, temperature and PM10 levels

Day No	Wind speed (KM/h)	Wind Dir. (in °)	Wind direction	Temp. (°C)	PM 10 (μg/m^3)	Actual target label	Predicted labels
50	18.3	126.69	SE	35.9	220.79	1	1
51	25.9	132.16	SE	33	161.92	1	1
52	20.6	126.82	SE	30.4	100.46	1	0
53	14.8	134.56	SE	32.1	112.59	1	1
54	14.8	164.5	SSE	31.6	160.36	1	1
55	9.4	153.48	SSE	29.8	191.47	1	0
56	11.1	129.31	SE	26.5	72.24	0	0
57	22.2	223.37	SW	31.5	148.76	1	1
58	29.4	225.32	SW	32.7	171.72	1	1
59	20.6	182.13	S	34.4	206	1	1
64	11.1	233.48	SW	23.1	88.21	0	1
103	14.8	163	SSE	25.5	42.5	0	0
110	11.5	155.7	SSW	27.4	94	0	0
142	13.3	219.67	SE	19.7	45	0	0
212	7.6	198.47	SSW	19.3	357.69	1	1
213	11.1	228.53	SW	17.9	423.06	1	1
214	18.3	246.49	WSW	17.6	432.28	1	1
215	14.8	251.09	WSW	17.7	506.44	1	1
216	7.6	222.65	SW	17.4	509.14	1	1
217	9.4	213.22	SSW	17.2	555.18	1	1

Multilayer Perceptron gave the best accuracy of 98.1%. The results are achieved after the removal of all the outliers. Therefore, we can say that MLP is the best technique among the three for the prediction of the PM 10 values.

These results, which were obtained on MLP, were achieved after a number of explorative experiments in which the various network parameters were varied to get their optimal values.

Table 2 does not show the entire dataset. Instead it contains a small sample describing only the highly positively and a few highly negatively correlated factors with PM 10 levels, their actual class labels and their corresponding predicted class labels. The high quality of the prediction of PM 10 levels which the model is successfully able to perform can be observed. As per the regulations stated in [7] the PM 10 levels above 100 $\mu g/m^3$ is considered as high (1) and levels below 100 $\mu g/m^3$ are considered low (0). It can be seen from the table that, out of the 20 selected instances the model is able to correctly predict 17 instances while only 03 are incorrectly predicted in this small sample space.

6 Conclusion

In this work, the task of constructing a pollution prediction model for Delhi was successfully accomplished. Of the various machine learning techniques used, we observed that Multilayer Perceptron gave the best results of all, with an overall accuracy of 98%.

Also, for the analysis part we found that out of all the meteorological factors—Wind Direction has the maximum positive correlation with PM 10 out of all the input factors considered individually. Temperature has the maximum negative correlation with PM 10 out of all the input factors considered individually. This means that if value of Temperature decreases then the value of PM 10 will increase and vice versa. Using this newly constructed model, highly accurate results can be predicted based on the current trends of the meteorological data, which can be, used abetment of particulate pollution in that area and help to develop pollution control strategies.

However this is a work attempts to identify a correlation between various metrological data and PM10 values, but doesn't consider vehicular traffic data. A further work can be seen as relevant where the vehicular traffic data may also be considered and correlation must be established between them and the particulate pollution. Also as the dataset used in this work is of about just over a year, hence the use of a dataset of a larger time span will help to establish the correlations and predictions better. A larger time span would mean repetition of weathers conditions which would help in asserting the reasons for pollution level changes at a certain time in the year.

References

1. World Health Organization. http://www.who.int/phe/health_topics/outdoorair/databases/cities/en/
2. Li, Y., Wang, W., Wang, J., Zhang, X., Lin, W., Yang, Y.: Impact of air pollution control measures and weather conditions on asthma during the 2008 Summer Olympic Games in Beijing. Int. J. Biometeorol. **55**(4), 547–554 (2011)
3. Wei, D.: Predicting Air Pollution Level in a Specific City. Stanford Publication (2014)
4. Pandey, G., Zhang, B., Jian, L.: Predicting submicron air pollution indicators: a machine learning approach. Environ. Sci.: Process. Impacts **15**(5), 996–1005 (2013)
5. Krupa, S., Nosal, M., Ferdinand, J.A., Stevenson, R.E., Skelly, J.M.: A multi-variate statistical model integrating passive sampler and meteorology data to predict the frequency distributions of hourly ambient ozone (O 3) concentrations. Environ. Pollut. **124**(1), 173–178 (2003)
6. Central Pollution Control Board. http://www.cpcb.gov.in
7. Central Pollution Control Board. http://cpcb.nic.in/National_Ambient_Air_Quality_Standards.php
8. Slini, T., Karatzas, K., Moussiopoulos, N.: Correlation of air pollution and meteorological data using neural networks. Int. J. Environ. Pollut. **20**(1–6), 218–229 (2003)