IJCSN - International Journal of Computer Science and Network, Volume 7, Issue 1, January 2018
ISSN (Online) : 2277-5420
www.IJCSN.org
Impact Factor: 1.5

Air Data Analysis for Predicting Health Risks

[1] Ranjana Gore; [2] Deepa Deshpande

[1] Computer Science & Engineering Department, Marathwada Institute of Technology, Aurangabad, Maharashtra 431001, Maharashtra

[2] Computer Science & Engineering Department, Jawaharlal Nehru Engineering College, Aurangabad, Maharashtra 431001, Maharashtra

Abstract - Air pollution is of much concern for the society. There are advancements in the technology leading to developed life. On the other hand it is responsible for green house gas emissions. There are different sources of GHGs like industries, agriculture construction sites, etc. GHGs are Nitrous Oxides, Sulphur Dioxide, Carbon Oxides, Methane, CFC and O3. These green house gases are the pollutants which hampers the quality of air. GHGs are responsible for global warming due to which there is ozone layer depletion. These pollutants may cause various health problems. Air quality can be assessed based on the Air quality levels (AQL). Air quality index can be obtained through different sensors or monitoring stations based on which air pollution related health concerns can be predicted. In this paper analysis was done on the dataset containing AQI of air pollutants such as NO2, O3, CO and SO2. The Random Forest algorithm shows accuracy of 93.467% while the Multiclass classifier algorithm shows the accuracy of 94.61%. The results shown that Multiclass classifier is better than the Random Forest algorithm. .

Keywords - *Data Mining; Random Forest; Multiclass Classifier; Air Quality Index.*

1. Introduction

Air pollutants are responsible for meticulous air pollution which hampers the human life. Air pollution may cause severe problems of respiratory system, skin problems and other health related issues, also affect eyes causing irritation. Therefore monitoring and analyzing air is a critical issue to have good and healthy life. This paper proposes the data mining techniques for analyzing the air pollution so that appropriate actions can be taken to lessen the adverse effects of increasing air pollution. An *air quality index* (AQI) is a numerical value which tells how polluted the air is. AQI levels are high, moderate, low or good etc. The increasing AQI value indicates that the air pollution is increasing. That is more the AQI values, more the air pollution. Smart city can be developed with a low-carbon in a sustainable way. Sustainability deals with economy, society and environment together. For this it is necessary to devise the action plan which can be done with analysis of pollutant concentrations and their increasing values.

2. Literature Review

The ambient air quality was assessed for Chennai on the data collected from CPCB. Data mining, Artificial Neural Network techniques were used [1]. The result obtained here would be used as an important guide for the Government and policy makers for developing new environmental and sustainable policies.

[2] worked on the relationship between air pollutants and admittance of patients. It has done the analysis for increase in the disease rate in the hospitals. The medical data and air pollutants data was collected for city Dhaka, Bangladesh. This paper uses k-means clustering algorithm for grouping the air pollutants while patient were classified using CART method.

The k-nearest neighbour technique was used to predict the value of air quality index [3]. This dataset contains sulphur dioxide, carbon monoxide, ozone and nitrogen oxide AQI values. The result has shown 0.6696% of Root mean squared error, 64.57% of relative absolute error and 87.10% of root relative squared error.

The dataset was analyzed by using Naïve Bayes algorithm and J48 algorithm. With Naïve Bayes 86.66% of accuracy was obtained while with J48 decision tree algorithm 91.99% of accuracy was obtained. The J48 algorithm gives better results than Naïve Bayes algorithm. These results can be more enhanced to achieve better accuracy [4].
In the proposed system Random forest algorithm and Multiclass classifier algorithms are selected for classification. Each record gets assigned to one and only

one class, this feature makes Multiclass classifier very efficient. Random Forest delivers better accuracy of 93.467% over J48 and Naïve Bayes Algorithms. Also the Multiclass classifier is more accurate than the random forest algorithm for this dataset.

3. Methodology

Data mining is defined as extraction of useful information from large dataset. Data can be collected from various data sources. There are two major tasks involved with Data Mining viz prediction and classification. Classification is the supervised learning technique where we can train the classifier for specific class labels. A new record provided to the classifier as input, corresponding class label gets assigned to that in the testing phase.

Classification - It can predict the class label for a new record to which the classifier is not trained.

Prediction - It is used to predict missing data values instead of class labels.

3.1 Random Forest Algorithm

Random Forest algorithm can be used for classification problem. This algorithm creates forest with number of trees. It gives highest accuracy with higher number of trees. In this technique the randomization is present in two ways; first random sampling of data for bootstrap samples as of bagging and another random selection for input attributes for constructing individual base decision trees [5].

3.2 Multiclass Classifier

The Multiclass classifier performs classification task with more than two classes. Multiclass classification makes the assumption that each sample is assigned to one and only one label.

3.3 Preprocessing

The readings for air quality index of air pollutants CO, SO2, NO2 and O3 are available in the dataset. There are some missing values for these air pollutants. These missing values can be handled by preprocessing. In this dataset different readings of AQI are available for each day.

Fig. 1 Model for Classification

The Air Quality Index is the maximum AQI value from the set of AQIs of NO2, SO2, O3 and CO in (2).

$$AQI = \text{Max}(NO2, SO2, O3, CO) \quad (2)$$

Fig. 1 shows model for classifying health risks based on the AQI values. The classifier is trained as follows: For AQI value in the range 0 to 50, the level of health concern is “GOOD”, for 51 to 100 the level of health concern is “MODERATE”, in the range of 101 to 150 “unhealthy for sensitive groups”, for the range 151 to 200 it is “UNHEALTHY”, in the range 201 to 300 “VERY UNHEALTHY” and above 300, the AQI level is hazardous.

4. Results

4.1 Random Forest

The total number of instances chosen for this study was 1837. The accuracy obtained for RF classifier is 93.467%. But this classifier was unable to correctly classify 120 records leading to an error rate of 6.53%.

Table 1: Confusion Matrix for RF Algorithm

a b c d e	*Class*
330 37 0 0 0	a = good
36 322 0 1 0	b = moderate
0 0 324 12 0	c= unhealthy_s
0 0 13 459 7	d =unhealthy
0 0 0 14 282	e = very_unhealthy

Table 1 shows the confusion matrix for random forest algorithm.

4.2 Multiclass Classifier

The Multiclass classifier has given the accuracy of 94.61% over 1837 samples. The error rate for the same is 5.38%.

Table 2: Confusion Matrix for MC Algorithm

a b c d e	*Class*
343 23 1 0 0	a = good
12 346 1 0 0	b = moderate

0 0 16 20 0	c= unhealthy_s
0 0 0 20 276	d =unhealthy
0 0 8 457 14	e = very_unhealthy

Table 2 shows the confusion matrix for Multiclass classifier algorithm.

4.3 Performance Comparison

After training, a new sample is tested whether class label assigned to it is correct or not. The Classifier accuracy can be obtained from (1) by using true positive (TP), true negative (TN), false positive (FP) and false negative (FN).

Accuracy= (TP + TN) / (FP + FN) …. (1)

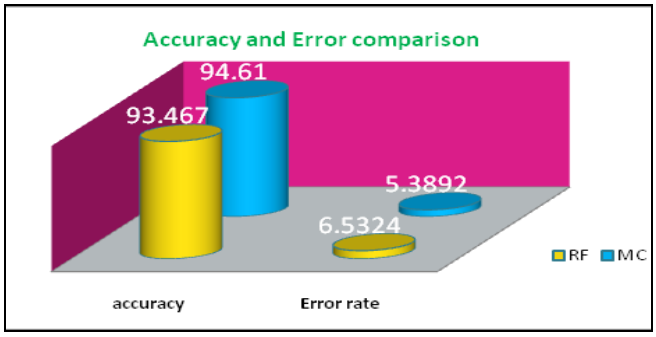

Fig. 2 Comparison of accuracy and error rate

$$Accuracy\ of\ RF = \frac{1717}{1837} = 93.46\%$$

$$Accuracy\ of MC = \frac{1738}{1837} = 94.61\%$$

$$Error\ rate\ of\ RF = \frac{120}{1837} = 6.53\%$$

$$Error\ Rate\ of\ MC = \frac{99}{1837} = 5.38\%$$

The Multiclass classifier is more accurate than the random forest algorithm in predicting the class label based on the given values of air pollutants SO2, NO2, CO and O3 as shown in fig. 2.

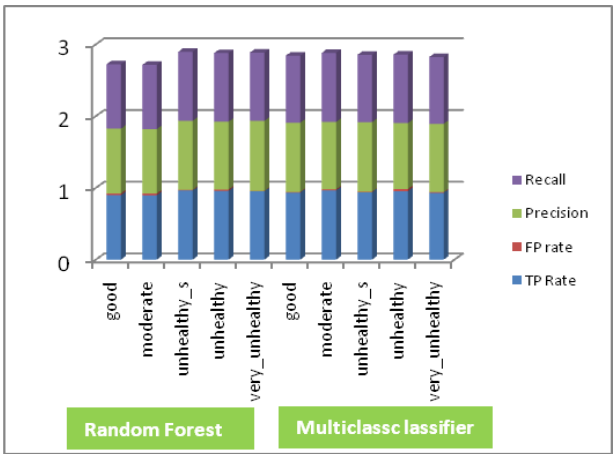

Fig. 3 Performance comparison of RF and MC algorithm

The dataset is evaluated for the MC and RF algorithms using precision, recall, TP rate and FP rate. The resultant graph is shown in figure 3 and the corresponding values are as shown in table 3.

Table 3: Performance measures for RF and MC algorithm

	Class	TP Rate	FP rate	Precision	Recall
RF	Good	0.899	0.024	0.9	0.899
	Moderate	0.897	0.025	0.9	0.897
	unhealthy_s	0.964	0.009	0.96	0.964
	Unhealthy	0.958	0.02	0.94	0.958
	very_unhealthy	0.953	0.005	0.98	0.953
MC	Good	0.935	0.008	0.97	0.935
	Moderate	0.964	0.016	0.94	0.964
	unhealthy_s	0.94	0.007	0.97	0.94
	Unhealthy	0.954	0.029	0.92	0.954
	very_unhealthy	0.932	0.009	0.95	0.932

4. Conclusions

It is very important to analyze the air quality to have good quality of life. This analysis plays very important role to develop smart city and to devise environmental policies. The data collected was analyzed with Multiclass label classifier and Random Forest classification techniques. The RF algorithm accuracy is 93.467% and accuracy of Multiclass label algorithm is 94.61%. The Multiclass label algorithm gives better accuracy than the RF classifier. The Root mean squared error is 0.1461, the relative absolute

error is 13.8137 % and the root relative squared error is 36.6507 %.

Acknowledgments

I would like to express my special thanks of gratitude to my teacher and those who have helped me in doing this research.

References

[1] https://www.kaggle.com/sogun3/uspollution/version/1 15-10-1016

[2] Ruhul Amin Dicken, "Analysis and classification of respiratory health risk with respect to air pollution levels", IEEE, 2015, pp. 1-6.

[3] Elia Georgiana Dragomir, "Air Quality Index Prediction using K-Nearest Neighbor Technique", Seria Matematica - Informatica – Fizica, Vol. LXII No. 1,2010, pp. 103 – 108

[4] Ranjana Waman Gore, Deepa S. Deshpande, "An Approach for Classification of Health Risks Based on Air Quality Levels", International Conference on Intelligent Systems and Information Management(ICISIM), 2017, IEEE.

[5] Leo Brieman, "Random Forests", Machine Learning, vol. 45, 2001, pp. 5-32.

[6] Carmen Capilla, "Neural networks data mining in an air quality database", International Environmental Modelling and SoftwareSociety (iEMSs)8th International Congress on Environmental Modelling and Software Toulouse, France, Sabin,e Sauvage, José-Miguel Sánchez-Pérez, Andrea Rizzoli (Eds.), 2016, pp 1279-1286.

[7] Krzysztof Siwek A, Stanisław Osowski, "Data Mining Methods For Prediction Of Air Pollution", International Journal Applied Mathathematics Computer Science, 2016, Vol. 26, No. 2, pp. 467–478.

[8] 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 ISSN: 2321-8169 Vol. 4 Issue: 2, pp. 153 – 157.

[9] Han J. and Kamber M. Data Mining: Concepts and techniques Morgan Kaufmann Publishers (2001).

[10] Sheng-Tun Lia, Li-Yen Shueb, " Data mining to aid policy making in air pollution management", Expert Systems with Applications,2004, vol. 27, pp. 331-340.

[11] Kavi K. Khedo, Rajiv Perseedoss and Avinash Mungur, "A wireless sensor network air pollution monitoring system", International journal of Wireless and mobile network, 2010, Vol 2, issue 2.

[12] Ioannis N. Athanasiadis, Kostas D. Karatzas and Pericles A. Mitkas. "Classification techniques for air quality forecasting." Fifth ECAI Workshop on Binding Environmental Sciences and Artificial Intelligence, 17th European Conference on Artificial Intelligence, Riva del Garda, Italy, August 2006.

[13] Pandey, Gaurav, Bin Zhang, and Le Jian. "Predicting submicron air pollution indicators: a machine learning approach." Environmental Science: Processes & Impacts 15.5 (2013): 996-1005.Elsevier, 2004, pp. 331–340.

First Author Ranjana Waman Gore did her BE (Computer Science & Engineering) from BAMU University in 2006. She did her post graduation in 2011 from GECA. She has worked as lecturer in Government College of Engineering, Aurangabad. She has worked at SPWEC College as head of the department and Assistant Professor. Currently she is working as assistant professor at Marathwada Institute of Technology, Aurangabad. She published two papers in IEEE International conferences, four papers in International Journals.

Second Author Deepa S. Deshpande did her BE(Computer) from Pune University in 1995. She did her MTech(Computer) from Pune university in 2006. She has completed her PhD from SRTMU in 2015. She has total 23 publications at national/International level. She is having 21 years of teaching experience and 1 year of industrial experience.