Comparison of Predictive Models in Data Mining and Impacts of Air Pollution in Metropolitan Cities

ASHA N
Research Scholar
Mother Teresa Women's University
Kodaikanal
ashan.gsc@gmail.com

Dr M P Indira Gandhi
Department of Computer Science
Mother Teresa Women's University
Kodaikanal
mpyazhini@gmail.com

***Abstract*— the prodigious increase in vehicle population and industries has led to the concentration of air pollutants in major metropolitan cities. The behavior of the air pollutants has a censorious slam on human health and environment. The health hazards caused due to alarming rate of air pollutants in large scale have to be forecasted and predicted to protect human health. Monitoring and predicting the voluminous data collected abundantly from various monitoring stations at urban area is opened challenge for discussion. This led to the scientist to look for several predicting data mining techniques and big data analytics to monitor and predict the urban air quality. The objective of this paper discusses about different approaches and comparative analysis of data mining techniques like linear regression, back propagation and big data analytics like Map reducing, Geostatistical methods to predict urban air quality.**

Keywords— air quality, big data analytics, data mining techniques

I. Introduction

Data mining, known as knowledge discovery in databases (KDD) is the process of discovering useful Knowledge from large amount of data stored in databases, data warehouses, or other information repositories. Data
Understanding starts with data collection and proceeds with activities to identify data quality problems, and to discover missing values into the data. Data preparation constructs the data to be modeled from the collected data. The modeling phase applies various modeling techniques, and determines the optimal values for parameters in models. The evaluation phase evaluates the model for the problem requirements [1]. Big data that concerns a large volume complex and growing data sets with several autonomous sources.Big data is rapidly growing in all science and engineering .Big data is characterized as (vvvvc) (i.e.) volume, velocity, variable, variety and complexity. In the recent years big data is growing in all science and engineering [7].Spatial interpolation is a study of environmental variables; the interpolation is used to study the spatial patterns of a phenomenon by estimating/predicting values at unsampled locations based on measurements at sample points [3]. Data mining Techniques and big data analytics are used to identify national air quality distribution whose continuous data emitted by different air quality monitoring stations at metropolitan cities that contains major composition of air pollutants *viz*., Sulphur Dioxide (SO2), Oxides of Nitrogen as NO2, Suspended Particulate Matter (SPM) and Respirable Suspended Particulate Matter (RSPM / PM10) have to be identified for regular monitoring at all the locations. The monitoring of meteorological parameters such as wind speed and wind direction, relative humidity (RH) and temperature should be also integrated with the monitoring of air quality.

II. Related Work

"Data mining in the Prediction of Impacts of Ambient Air Quality Data Analysis in Urban and Industrial Area " by S. Christy and Dr. V. Khanaa.This paper describes the Air quality is monitored by air quality monitoring stations in Chennai through the use of wireless sensors deployed in huge numbers around the city and industrial areas. 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, Feed Forward Neural Networks and Multilayer Perceptron neural network models are used to predict the maximum concentration of the pollutants in Chennai that provides the policy makers in contriving the future air pollution standard policies.

"Hadoop-based distributed system for online prediction of air pollution based on support vector machine" byZ. Ghaemia, M. Farnaghib, A. Alimohammadib.The purpose of this paper is to present an online forecasting approach based on Support Vector Machine (SVM) to predict the air quality one day in advance. In order to overcome the computational requirements for large-scale data analysis, distributed computing based on the Hadoop platform has been employed to leverage the processing power of multiple processing units. The Map Reduce programming model is adopted for massive parallel processing in this study.The paperfocuses on big data analytics techniques for online forecasting to predict the air pollution of Tehran. The results have proved best processing time and efficiency to deal large scale air pollution prediction problems.

"Predictive mapping of air pollution involving sparse spatial observations" by Jeremy E. Diema,*, Andrew C. Comrieba.This paper's main objective is to outline an approach that others can use to map air pollution concentrations for areas with a limited number of spatial observations, but which have an abundance of temporal observations and sufficient ancillary geospatial data. The paper illustrate about geostatistical methods are used to predict ozone level concentration of pollutants that varies with planetary boundary layer such as temperature, wind speed, and wind direction.

"Data mining methods for prediction of air Pollution" by Krzysztof SIWEK and Stanislaw OSOWSKI.The paper discusses the methods of data mining for prediction of air pollution. Two problems in such prediction are important: the generation and selection of the prognostic features, and final prognosis of the pollution level for the next day on the basis of the data of the previous day. This paper analyzes and compares two methods of feature selection like genetic algorithm, and the linear method of stepwise fit. The factors that influence the predictions are selected as features and compared with Multilayer perceptron and Support Vector machine.

"Evaluation of optimum methods for predicting pollution concentration in GIS environment" by R. Shad, H Ashoori, N. Afshari.Analysis to create predicted surface for unmeasured points in study area. For This purpose, Ground stations and MODIS image of Tehran are used for collecting online air pollution information. Then, different geostatistical methods have been used for finding out the optimum prediction method for air pollution, based on received observations. One of the important spatial analyses for this application in GIS environment is surface simulation using Geostatistical methods used in GIS models for optimum decision making. The ArcGIS application developed by ESRI was selected for analyzing different collected data because of its dynamic Geostatistical environmental models. In this software localization of information extracted of MODIS image, industries, emission patterns, etc can be displayed together. GIS for optimum decision making based on the air quality factors which can be collected as maps, satellite images, and ground stations data. Geostatistical methods are urgently needed for the amount of pollution in everywhere.

"A multi-agent framework for a hadoop based air quality decision support system" byAbdelaziz E Fazziki , AbderrahmaneSadiq , Jamal Ouarzazi , Mohamed Sadgal .
Approach based on the use of the agent technology and big data concept. For the air quality data collection and analysis. Air quality management support system for the Marrakech city is presented. HBase for data storage and a Map Reduce based forecasting process; artificial neural network (ANN) based prediction and K-means as clustering algorithm. Data are extracted and stored into a Hadoop HBase. HBase is a data-base with high reliability, high performance, column storage, scalable characteristics based on the Hadoop distributed file system (HDFS). For big number of small data blocks, the processing jobs increases the number of collaboration during the Map and Reduce operation as Hadoop has the advantage of handling large size of files.

III. 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 [9]. At 13 locations the ambient air quality monitoring is being carried out regularly twice a week, 24 hours a day using manual equipments for parameters such as PM10, SO2 and NO2 and at 2 locations monitoring is carried out round the clock using Continuous Ambient Air Quality Monitoring Stations (CAAQMS) for PM10 (RSPM), SO2, NO2 and CO. Vehicles are more threatening than other sources of pollution as they discharge emissions directly into the air. The most critical form of pollution in Bangalore is Respirable Suspended Particulate Matter (RSPM), according to the Karnataka State Pollution Control Board (KSPCB) [10]. The KSPCB maintains a website for publishing archived and real time pollutant information and forecasting.

For instance, the homogeneous regions could be varied when the scale of temporal data is changed from small scale (e.g., hourly, daily, etc.) to large scale (e.g., monthly, seasonally, or annually). The selection of an appropriate scale is dependent on the application purpose [1]. The data collected in large scale have to be predicted using proper predictive techniques to measure the maximum concentration of air pollutants affecting the human health. The selection of predictive technique should support real time analysis. Figure 1 demonstrates the study area and distribution of air pollution stations.

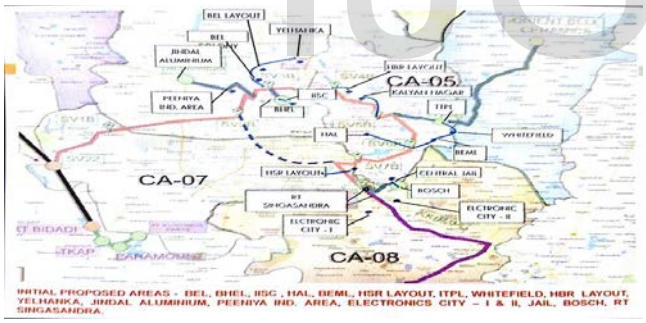

Figure 1: Air Quality monitoring stations at Bangalore city

IV. Health Impacts of Air Pollution

Air pollutants that are inhaled have serious impact on human health affecting the lungs and the respiratory system they are also taken up by the blood and pumped all round the body [11].Carbon monoxide (CO) is primarily emitted by motor vehicles and is dangerous when a motor vehicle is operating in confined space if emitted in large volumes. It reduces the oxygen circulation in the bloodstream by interaction with the hemoglobin in blood .Sulphur di oxide (SO2) the major sources are power plants, refineries and smelters, this as adverse impact on human health as it affects the respiratory system that results in asthma. RSPM (PM10)-Respirable suspended particle are smaller 10 mm in diameter are known as inhalable particulate matter, and those smaller than 2.5 mm (PM2.5) are called respirable particulate matter[8]. Exposure to high concentrations of PM_{10} can result in a number of health impacts ranging from coughing and wheezing to asthma attacks and bronchitis to high blood pressure, heart attack, strokes and premature death[12].

V. Comparative Analysis of the Predictive Techniques used in Literature Review

From the literature review, it could be understood that many techniques are used to predict air pollution which are summarized below in terms of its usage, type of data used, speed and accuracy.

1. **Domain of study:** Data Mining
- **Technique:** Linear regression and Multiple Regression model.
- **Usage:** Used as most basic statistical tool to make predictions known to humankind, with applications in statistics, finance, medicine, economics, and psychology.
- **Type of data used:** Continuous values.
- **Speed:** Linear regression works for one predictor variable, multiple regression model works for more than one Predictor variable. Its speed depends on number of predictor variables.
- **Accuracy:** There may be measurement error in the variables, model may be interpreted wrongly and cannot be used when data is very large.

2. **Domain of study:** Neural Network
- **Technique:**Artificial Neural Network (ANN) model, Feed Forward Neural Networks and Multilayer Perceptron neural network models.
- **Usage:** In the field of environmental engineering.
- **Type of data used:** It works with high tolerance of noisy data and as the ability to classify patterns on which they are not trained.
- **Speed:** Training the network for the given data sets requires time.
- **Accuracy:** The most accurate predictions are obtained using normalized variables.

3. **Domain of study:**Big data Analytics
- **Technique:** Map reducing model
- **Usage:** Map Reduce is a framework for processing parallelizable problems across large datasets using a large number of

computers (nodes), collectively referred to as a cluster

- **Type of data used:** Works with both structured and unstructured data
- **Speed:** Good memory usage and computation time is more, since map reducing is a parallel processing algorithm that works on one or more cluster.
- **Accuracy:** The system could not only achieve promising results, but also possess a good prediction performance as well.

4. **Domain of study:** Geostatistics
   - **Technique:**Geostatistics is concerned with a variety of techniques aimed at understanding and modeling spatial variability through prediction and simulation like co-Kriging interpolation.
   - **Usage:** The primary aim of geostatistics is to estimate the spatial relationship between sample values. The estimate is used to make spatial prediction of unobserved values from neighboring samples and to give an estimate of the variance of the prediction error.
   - **Type of data used:** Maps and satellite images.
   - **Speed:** Depends on temporal factors.
   - Co-Kriging interpolation is more accurate by **Accuracy:**producing and evaluating prediction standard error maps.

VI. Map Reducing model

Big Data is a collection of large datasets that cannot be processed using traditional computing techniques. For example, the volume of data Face book or YouTube need require it to collect and manage on a daily basis; can fall under the category of Big Data. However, Big Data is not only about scale and volume.

It also involves one or more of the following aspects − Velocity, Variety, Volume, and Complexity [13]. Map Reduce, introduced by Google provides parallel computing power to process parallelizable problems across huge datasets. Map Reduce based process to make needed algorithms applicable to large scale data and give a great flexibility and speed to execute a process over the distributed framework [2]. Each Map Reduce process consists of two phases: the Map phase and the Reduce phase.

- The Map task takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key-value pairs).

- The Reduce task takes the output from the Map as an input and combines those data tuples (key-value pairs) into a smaller set of tuples.The reduce task is always performed after the map job [13].

In other words, the reduce function accepts an intermediate key and a set of values for that key. It merges together these values to form a possibly smaller set of values. Finally, the reducer sends the obtained results to the output. A map/reduce job consists of some independent map tasks and some independent reduce tasks. The number of Mappers determines the level of parallelism. The input data is divided into some independent chunks which are processed by the map tasks in a parallel manner. The outputs of the Mappers are then sorted by the Map Reduce framework and sent to the reduce tasks as inputs. Both the input and the output of the job are stored in a file system [2].

VII. Conclusion

Since the data is continuously emitted by the monitoring station is enormous on an daily basis and have to be stored that requires lot of space and processed at high computing speed to make immediate decisions so as to protect human health and measures to reduce air pollution. To overcome the challenge of large scale data of air pollutants it is theoretically understood from the above comparative analysis of predictive techniques that Hadoop-Map Reduce best suits for prediction of air pollution as it supports large memory storage and high computing time and distributed parallel processing. The data set of each monitoring station can be split into data nodes and training each subset of data in parallel will be the future research work.

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. Z. Ghaemi, M. Farnaghib, A. Alimohamma "Hadoop-based distributed system for online prediction of air pollution based on support vector machine" The International Archives of the

Photogrammetric, Remote Sensing and Spatial Information Sciences Nov 2015.

3. 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.
4. Krzysztof Siwek, Stanislaw Osowski "Data mining methods for prediction of air pollution -Extended summary "Warsaw University of Technology, POLAND.
5. R. Shad a, H Ashoori b, N. Afshari b" Evaluation of optimum methods for predicting pollution concentration in Gis environment" A Faculty of Geodesy and Geomatics Eng, K.N.Toosi University of Technology, Roozbeh_Shad@yahoo.com .
6. Abdelaziz El Fazziki , AbderrahmaneSadiq , Jamal Ouarzazi , Mohamed Sadgal "A multi-agent framework for a Hadoop based air quality decision support system " Computer Systems Engineering Laboratory, Cadi Ayyad University of Marrakech.
7. Abinaya.K "Data Mining with Big Data e-Health Service Using Map Reduce"International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 2, February 2015.
8. Khaled Ahmed Ali Abdulla Ai Koas "Gis based mapping and Statistical Analysis of Air Pollution and Mortality in Brisbane Australia" Queens land University of technology, April 2010.
9. kspcb.kar.nic.in / The Karnataka State Pollution Control Board for Prevention and Control of Water Pollution constituted by the Government of Karnataka.
10. Article "Two-wheelers are biggest pollutants in Bangalore" Saswatimukherjee b| tnn | apr 14, 2013, 04.09 am ist.
11. edugreen.teri.res.in/explore/air/health.htm
12. http://www.marlborough.govt.nz/Environment/Air-Quality/Smoke-and--Smog/Health-effects-of-PM10.aspx
13. https://www.tutorialspoint.com/map_reduce/map_reduce_introduction.htm