IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Air Quality Index Using Artificial Intelligence And Machine Learning

Mr. Aishwary Suhas Shivarkar, Prof. Vaishali Bagade, Department of Electronics and telecommunication engineering, Alamuri Ratnamala Institute of Engineering and Technology, A.S. Rao Nagar, Vill-Sapgaon, Tal-Shahpur, Dist-Thane, University of Mumbai, Maharashtra, India.

ABSTRACT_ Air pollution and its prevention are constant scientific challenges during the last decades. However, they remain huge global problems. Affecting the human respiratory and cardiovascular system, they are a cause of increased mortality and increased risk for diseases for the population. Many efforts from both local and state governments are done to understand and predict the air quality index aiming improved public health. This paper is one scientific contribution toward this challenge. We compare four simple machine learning algorithms, Liner Regression, Lacco CV Regressor, support vector machines, and decision tree. The following dataset "India Air Quality Data" has been analyzed. The results are promising and it was proven that implementation of these algorithms could be very efficient in predicting air quality index.

KEYWORDS: AIR POLLUTION, MACHINE LEARNING, INTERNET OF THINGS (IOT), LINER REGRESSION, AIR QUALITY DATA.

I. Introduction

Artificial intelligence and machine learning are areas of the biggest rise in the last year. The science of artificial intelligence where the system decides on its own, instead of working only by orders given by programmers as traditional programming works, gradually started influencing all aspects of our life. Starting from early-stage startup companies and ending to large platform vendors, for all of them, artificial intelligence and its part machine learning have become the key focus areas.

Machine learning is an area where a system that implements artificial intelligence gathers data from sensors in an environment and learns how to act. One of the reasons why we choose machine learning to predict the air quality index was the ability to adapt machine learning (ML) algorithms.

In this paper, three supervised learning algorithms Liner Regressor (LR), Support Vector Machines (SVM) and Decision Tree (DT), and Laso CV are compared.

Many researchers implement some of the algorithms we are using, such as NN [1], NN and fuzzy systems [2], SVM [3], SVM for regression [4], fuzzy logic [5], DT [6], k-NN [7], but none of them compare their performance as one research for all of them four at the same conditions and the same data.

Adverse health impacts from exposure to outdoor air pollutants are complicated functions of pollutant compositions and concentrations [1]. Major outdoor air pollutants in cities include ozone (O3), particle matter (PM), sulfur dioxide (SO2), carbon monoxide (CO), nitrogen oxides (NOx), volatile organic compounds (VOCs), pesticides, and metals, among others [2,3]. Increased mortality and morbidity rates have been found in association with increased air pollutants (such as O3, PM and SO2) concentrations [3–5]. According to the report from the American Lung Association [6], a 10 parts per billion (ppb) increase in the O3 mixing ratio might cause over 3700 premature deaths annually in the United States (U.S.). Chicago, as for many other megacities in U.S., has struggled with air pollution as a result of industrialization and urbanization. Although O3 precursor (such as VOCs, NOx, and CO) emissions have significantly decreased since the late 1970s, O3 levels in Chicago have not been in compliance with standards set by the Environmental Protection Agency (EPA) to protect public health [7]. Particle size is critical in determining the particle deposition location in the human respiratory system [8]. PM2.5, referring to particles with a diameter less than or equal to 2.5 um, has been an increasing concern, as these particles can be deposited into the lung gas-exchange region, the alveoli [9]. The U.S. EPA revised the annual standard of PM2.5 by lowering

the concentration to 12 ug/m3 to provide improved protection against health effects associated with long- and short-term exposure [10]. SO2, as an important precursor of new particle formation and particle growth, has also been found to be associated with respiratory diseases in many countries [11–15]. Therefore, we selected O3, PM2.5 and SO2 for testing in this study.

Meteorological conditions, including regional and synoptic meteorology, are critical in determining the air pollutant concentrations [16–21]. According to the study by Holloway et al. [22], the O3 concentration over Chicago was found to be most sensitive to air temperature, wind speed and direction, relative humidity, incoming solar radiation, and cloud cover. For example, a lower ambient temperature and incoming solar radiation slow down photochemical reactions and lead to less secondary air pollutants, such as O3 [23]. Increasing wind speed could either increase or decrease the air pollutant concentrations. For instance, when the wind speed was low (weak dispersion/ventilation), the pollutants associated with traffic were found at the highest concentrations [24,25]. However, strong wind speeds might form dust storms by blowing up the particles on the ground [26]. High humidity is usually associated with high concentrations of certain air pollutants (such as PM, CO and SO2) but with low concentrations of other air pollutants (such as NO2 and O3) because of various formation and removal mechanisms [25]. In addition, high humidity can be an indicator of precipitation events, which result in strong wet deposition leading to low concentrations of air pollutants [27]. Because various particle compositions and their interactions with light were found to be the most important factors in attenuating visibility [28,29], low visibility could be an indicator of high PM concentrations. Cloud can scatter and absorb solar radiation, which is significant for the formation of some air pollutants (e.g., O3) [23,30]. Therefore, these important meteorological variables were selected to predict air pollutant concentrations in this study. Statistical models have been applied for air pollution prediction on the basis of meteorological data [31–35]. However, existing studies on statistical modeling have mostly been restricted to simply utilizing standard classification or regression models, which have neglected the nature of the problem itself or ignored the correlation between sub-models in different time slots. On the other hand, machine learning approaches have been developing for over 60 years and have achieved tremendous success in a variety of areas [36–41].

There exist various new tools and techniques invented by the machine learning community, which allow for more refined modeling of a specific problem. In particular, model regularization is a fundamental technique for improving the generalization performance of a predictive model. Accordingly, many efficient optimization algorithms have been developed for solving various machine learning formulations with different regularizations in this study, we focus on refined modeling for predicting hourly air pollutant concentrations on the basis of historical metrological data and air pollution data. A striking difference between this work and the previous works is that we emphasize how to regularize the model in order to improve its generalization performance and how to learn a complex regularized model from big data with advanced optimization algorithms. We collected 10 years' worth of meteorological and air pollution data from the Chicago area. The air pollutant data was from the EPA [42,43], and the meteorological data was from MesoWest [44]. From their databases, we fetched consecutive hourly measurements of various meteorological variables and pollutants reported by two air quality monitoring stations and two air pollutant monitoring sites in the Chicago area. Each record of hourly measurements included meteorological variables such as solar radiation, wind direction and speed, temperature, and atmospheric pressure; as well as air pollutants, including PM2.5, O3, and SO2. We used two methods for model regularization: (i) explicitly controlling the number of parameters in the model; (ii) explicitly enforcing a certain structure in the model parameters. For controlling the number of parameters in the model, we compared three different model formulations, which can be considered in a unified multi-task learning (MTL) framework with a diagonal- or full-matrix model. For enforcing the model matrix into a certain structure, we have considered the relationship between prediction models of different hours and compared three different regularizations with standard Frobenius norm regularization. The experimental results show that the model with the intermediate size and the proposed regularization, which enforces the prediction models of two consecutive hours to be close, achieved the best results and was far better than standard regression models. We have also developed efficient optimization algorithms for solving different formulations and demonstrated their effectiveness through experiments. The rest of the paper is organized as follows. In Section 2, we discuss related work. In Section 3, we describe the data collection and preprocessing. In Section 4, we describe the proposed solutions, including

formulations, regularizations and optimizations. In Section 5, we present the experimental studies and the results. In Section 6, we give conclusions and indicate future work.

II. Related Work

Many previous works have been proposed to apply machine learning algorithms to air quality predictions. Some researchers have aimed to predict targets into discretized levels. Kalapanidas et al. [32] elaborated effects on air pollution only from meteorological features such as temperature, wind, precipitation, solar radiation, and humidity and classified air pollution into different levels (low, med, high, and alarm) by using a lazy learning approach, the case-based reasoning (CBR) system. Athanasiadis et al. [45] employed the s-fuzzy lattice neurocomputing classifier to predict and categorize O 3 concentrations into three levels (low, mid, and high) on the basis of meteorological features and other pollutants such as SO2, NO, NO2, and so on. Kurt and Oktay [33] modeled geographic connections into a neural network model and predicted daily concentration levels of SO2, CO, and PM10 3 days in advance. However, the process of converting regression tasks to classification tasks is problematic, as it ignores the magnitude of the numeric data and consequently is inaccurate. Other researchers have worked on predicting concentrations of pollutants. Corani [46] worked on training neural network models to predict hourly O3 and PM10 concentrations on the basis of data from the previous day. Mainly compared were the performances of feed-forward neural networks (FFNNs) and pruned neural networks (PNNs). Further efforts have been made on FFNNs: Fu et al. [47] applied a rolling mechanism and gray model to improve traditional FFNN models. Jiang et al. [48] explored

multiple models (physical and chemical model, regression model, and multiple layer perceptron) on the air pollutant prediction task, and their results show that statistical models are competitive with the classical physical and chemical models. Ni, X. Y. et al. [49] compared multiple statistical models on the basis of PM2.5 data around Beijing, and their results implied that linear regression models can in some cases be better than the other models. MTL focuses on learning multiple tasks that have commonalities [50] that can improve the efficiency and accuracy of the models. It has achieved tremendous successes in many fields, such as natural language processing [37], image recognition [38], bioinformatics [39,40], marketing prediction [41], and so on. A variety of regularizations can be utilized to enhance the commonalities of the related tasks, including the `2,1-norm [51], nuclear norm [52], spectral norm [53], Frobenius norm [54], and so on. However, most of the former machine learning works on air pollutant prediction did not consider the similarities between the models and only focused on improving the model performance for a single task, that is, improving prediction performance for each hour either separately or identically.

Therefore, we decided to use meteorological and pollutant data to perform predictions of hourly concentrations on the basis of linear models. In this work, we focused on three different prediction model formulations and used the MTL framework with different regularizations. To the best of our knowledge, this is the first work that has utilized MTL for the air pollutant prediction task. We exploited analytical approaches and optimization techniques to obtain the optimal solutions. The model's evaluation metric was the root-mean-squared error (RMSE).

III. Methodology

There are two primary phases in the system: 1. Training phase: The system is trained by using the data in the data set and fits a model (line/curve) based on the algorithm chosen accordingly. 2. Testing phase: the system is provided with the inputs and is tested for its working. The accuracy is checked. And therefore, the data that is used to train the model or test it, has to be appropriate. The system is designed to detect and predict AQI levels and hence appropriate algorithms must be used to do the two different tasks. Before the algorithms are selected for further use, different algorithms were compared for their accuracy. The well-suited one for the task was chosen.

IV. Data preprocessing and normalization

a) Data Source

To predict the air quality index of a particular region, we need the pollutant concentration of all the gases which will be available in the cpcb.nic.in the website, which holds all the data that pollutes the cities every year. The AQI formulae will be applied to calculate the AQI. Datasets will be imported inside the directory and null values will be set to the infinite data. The predicted and actual values will be represented using the Bar-graph analysis to remove the outliers.

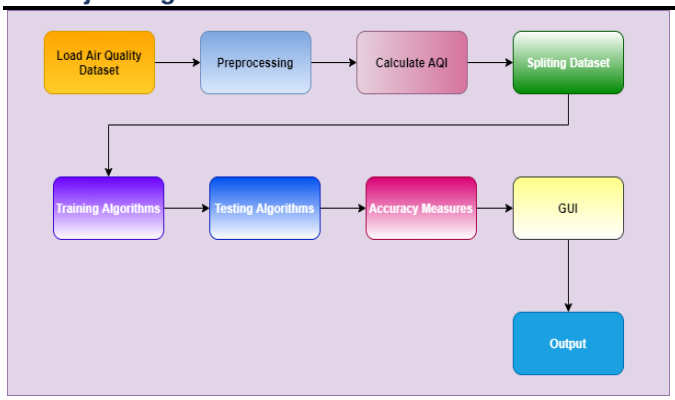

Fig 1: Flow Diagram

b) AQI Calculation

The AQI is an index for reporting daily air quality. It tells you how clean or polluted your air is, and what associated health effects might be a concern for you. The AQI focuses on health affects you may experience within a few hours or days after breathing polluted air. EPA calculates the AQI for five major air pollutants regulated by the Clean Air Act: ground-level ozone, particle pollution Air quality directly affects (also known as particulate quality of life. matter), carbon monoxide, sulfur dioxide, and nitrogen dioxide. For each of these pollutants, EPA has established national air quality standards to protect public health.

$$AQI = AQI_{min} + \frac{PM_{Obs} - PM_{Min}}{AQI_{Max} - AQI_{Min}}(PM_{Max} - PM_{Min})$$

Fig 2: AQI formula

AQI is calculated in the range of 0-500, we are scaling the values according to the AQI calculation formula

The index category for SO2 is scaled between 0-1600. So on applying the formula which is used to calculate AQI

The index category for NO2 is scaled between 0-400. So on applying the formula which is used to calculate AQI

The index category for rspm is scaled between 0-400. So on applying the formula which is used to calculate AQI

The index category for rspm is scaled between 0-430. So on applying the formula which is used to calculate AQI

The index category for rspm is scaled between 0-430. So on applying the formula which is used to calculate AQI

The purpose of the AQI is to understand what local air quality means to your health. Also, it is scaled from 0 to 500.

V. Proposed System

Step 1: Extraction of the historical dataset.
Step 2: Data preprocessing and normalization.
Step 3: Training ad TestingModel
Step 4: Algorithms
Step 5: Details of Hardware & Software with GIU

1. Extraction of the historical dataset

a) Missing values being filled in columns

Since we already know that our dataset contains missing values, we need to fill them for our further analysis. We will be using Imputation to fill in our missing values. Imputation is the process of replacing missing data with substituted values. Because missing data can create problems for analyzing data, imputation is seen as a way to avoid pitfalls involved with listwise deletion of cases that have missing values.

b) Understanding the pollutants briefly.

NO2: Nitrogen Dioxide and is emitted mostly from combustion from power sources or transport.

SO2: Sulphur Dioxide and is emitted mostly from coal burning, oil burning, and manufacturing of Sulphuric acid.

spm: Suspended particulate matter and is known to be the deadliest form of air pollution. They are microscopic in nature and are found to be suspended in the earth's atmosphere.

rspm: Respirable suspended particulate matter. A subform of spm and are responsible for respiratory diseases.

pm2_5: Suspended particulate matter with diameters less than 2.5 micrometers. They tend to remain suspended for longer durations and are potentially very harmful.

VI. CONCLUSIONS

We have successfully performed a comparative study of the algorithms for this thesis. SVR has the best results as compared to the other algorithms. The linear Regression algorithm has the worst accuracy. Thus, by performing a comparative study of algorithms we have successfully boosted the overall accuracy of the system. In the future, more complex or hybrid-based boosting algorithms can be used for obtaining higher accuracy.

References

1. Curtis, L.; Rea, W.; Smith-Willis, P.; Fenyves, E.; Pan, Y. Adverse health effects of outdoor air pollutants. Environ. Int. 2006, 32, 815–830.
2. Mayer, H. Air pollution in cities. Atmos. Environ. 1999, 33, 4029–4037.
3. Samet, J.M.; Zeger, S.L.; Dominici, F.; Curriero, F.; Coursac, I.; Dockery, D.W.; Schwartz, J.; Zanobetti, A. The national morbidity, mortality, and air pollution study. Part II: Morbidity and mortality from air pollution in the United States. Res. Rep. Health Eff. Inst. 2000, 94, 5–79.
4. Dockery, D.W.; Schwartz, J.; Spengler, J.D. Air pollution and daily mortality: Associations with particulates and acid aerosols. Environ. Res. 1992, 59, 362–373.
5. Schwartz, J.; Dockery, D.W. Increased mortality in Philadelphia associated with daily air pollution concentrations. Am. Rev. Respir. Dis. 1992, 145, 600–604.
6. American Lung Association. State of the Air Report; ALA: New York, NY, USA, 2007; pp. 19–27.
7. Environmental Protection Agency (EPA). Region 5: State Designations, as of September 18, 2009. Available online: https://archive.epa.gov/ozonedesignations/web/html /region5desig.html (accessed on 17 December 2017).
8. Hinds, W.C. Aerosol Technology: Properties, Behavior, and Measurement of Airborne Particles; John Wiley & Sons:
Hoboken, NJ, USA, 2012.
9. Soukup, J.M.; Becker, S. Human alveolar macrophage responses to air pollution particulates are associated with insoluble components of coarse material, including particulate endotoxin. Toxicol. Appl. Pharmacol. 2001, 171, 20–26.
10. Environmental Protection Agency (EPA). CFR Parts 50, 51, 52, 53, and 58-National Ambient Air Quality Standards for Particulate Matter: Final Rule. Fed. Regist. 2013, 78, 3086–3286.
11. Schwartz, J. Short term fluctuations in air pollution and hospital admissions of the elderly for respiratory
disease. Thorax 1995, 50, 531–538.
12. De Leon, A.P.; Anderson, H.R.; Bland, J.M.; Strachan, D.P.; Bower, J. Effects of air pollution on daily hospital admissions for respiratory disease in London between 1987-88 and 1991-92. J. Epidemiol. Community Health 1996, 50 (Suppl. 1), s63–s70.
13. Birmili, W.; Wiedensohler, A. New particle formation in the continental boundary layer: Meteorological and gas phase
parameter influence. Geophys. Res. Lett. 2000, 27, 3325–3328.
14. Lee, J.-T.; Kim, H.; Song, H.; Hong, Y.C.; Cho, Y.S.; Shin, S.Y.; Hyun, Y.J.; Kim, Y.S. Air pollution and asthma
among children in Seoul, Korea. Epidemiology 2002, 13, 481–484.
15. Cai, C.; Zhang, X.; Wang, K.; Zhang, Y.; Wang, L.; Zhang, Q.; Duan, F.; He, K.; Yu, S.-C. Incorporation of new particle formation and early growth treatments into WRF/Chem: Model improvement, evaluation, and impacts of anthropogenic aerosols over East Asia. Atmos. Environ. 2016, 124, 262–284.
16. Kalkstein, L.S.; Corrigan, P. A synoptic climatological approach for geographical analysis: Assessment of sulfur dioxide concentrations. Ann. Assoc. Am. Geogr. 1986, 76, 381–395.
17. Comrie, A.C. A synoptic climatology of rural ozone pollution at three forest sites in Pennsylvania. Atmos. Environ. 1994, 28, 1601–1614.
18. Eder, B.K.; Davis, J.M.; Bloomfield, P. An automated classification scheme designed to better elucidate the dependence of ozone on meteorology. J. Appl. Meteorol. 1994, 33, 1182–1199.
19. Zelenka, M.P. An analysis of the meteorological parameters affecting ambient concentrations of acid aerosols in Uniontown, Pennsylvania. Atmos. Environ. 1997, 31, 869–878.
20. Laakso, L.; Hussein, T.; Aarnio, P.; Komppula, M.; Hiltunen, V.; Viisanen, Y.; Kulmala, M. Diurnal and annual characteristics of particle mass and number concentrations in urban, rural and Arctic environments in Finland. Atmos. Environ. 2003, 37, 2629–2641.
21. Jacob, D.J.;Winner, D.A. Effect of climate change on air quality. Atmos. Environ. 2009, 43, 51–63.
22. Holloway, T.; Spak, S.N.; Barker, D.; Bretl, M.; Moberg, C.; Hayhoe, K.; Van Dorn, J.; Wuebbles, D. Change in ozone air pollution over Chicago associated with global climate change. J. Geophys. Res. Atmos. 2008, 113, doi:10.1029/2007JD009775.
23. Akbari, H. Shade trees reduce building energy use and CO2 emissions from power plants. Environ. Pollut. 2002, 116, S119–S126.
24. DeGaetano, A.T.; Doherty, O.M. Temporal, spatial and meteorological variations in hourly PM 2.5 concentration extremes in New York City. Atmos. Environ. 2004, 38, 1547–1558.
25. Elminir, H.K. Dependence of urban air pollutants on meteorology. Sci. Total Environ. 2005, 350, 225–237.
26. Natsagdorj, L.; Jugder, D.; Chung, Y.S. Analysis of dust storms observed in Mongolia during 1937–1999. Atmos. Environ. 2003, 37, 1401–1411.