IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue-15, May 2019 E-ISSN: 2347-2693

Development of Machine Learning-Based Predictive Models for Air Quality Analysis and Prediction

[1*]Ashok Shah, [2]PrayashRimal, [3]DevBhasker Singh, 4 Jagadeesh B N

[1]Dept. of CSE, East West Institute of Technology, Visvesvaraya Technological University, Bangalore, INDIA

DOI: https://doi.org/10.26438/ijcse/v7si15.3235 | Available online at: www.ijcseonline.org

***Abstract*—** One of the biggest environmental problems right now is air pollution. Air quality is needed to be consistently monitored and assessed to ensure better living conditions. The U.S. Environmental Protection Agency (EPA) uses the air quality index (AQI) to standardize the air quality. However, AQI requires precise and accurate sensor readings and complex calculations, making it not feasible for portable air quality monitoring devices. The aim of this paper is to find an alternative way of monitoring and characterizing air quality through the use of integrated gas sensors and building predictive models using machine learning algorithms that can be used to obtain data-driven solutions to mitigate the risk of air pollution.

***Keywords*—** AQI(Air Quality Index),Data Cleaning, Softmax Function

I. INTRODUCTION

Monitoring air quality is one of the best ways to prevent theharmful effects of air pollution. Having the informationabout the quality of air can lead to formulating suggestionsand datadriven recommendations to mitigate the possible harmful effects it can bring.

In this work, a business intelligent model has been developed, to predict the air quality, based on a specific business structure using a suitable machine learning technique. The model was evaluated by a scientific approach to measure accuracy. As we are using unlabeled data for Air quality analysis and prediction, so to build the classification model we are using Random Forest Classifier.

II. RELATED WORK

In this section, we discuss three main research categories related to our work, including air pollution control, semisupervised learning and feature selection.

Air pollution control: Various approaches have been proposed to apply data mining to the topics of prediction, and feature analysis for air pollutioncontrol in the recent literature. Different from most of the existing work which focus on one or two specific topics in the area of air pollution controlling, our work unify the interpolation, prediction, feature selection and analysis of fine-grained air quality into one model.

Semi-supervised learning: Semi-supervised learning is a class of supervised learning tasks and techniques that also make use of unlabelled data for training – typically a small amount of labelled data with a large amount of unlabelleddata. This characteristic is utilized in semi-supervised learning to better exploit the geometric structure of the spatiotemporal data, and to achieve the purpose of interpolation.

Feature selection: In many real-world applications, feature selection techniques have become an apparent need, give a review of feature selection techniques in bioinformatics.

III. METHODOLOGY

In this work, a business intelligent model has been developed, to predict the air quality, based on a specific business structure using a suitable machine learning technique. The model was evaluated by a scientific approach to measure accuracy. As we are using unlabelled data for Air quality analysis and prediction. To build the classification model we are using Random Forest Classifier.

It will cover the details explanation of methodology that is being used to make this project complete and working well. Many methodology or findings from this field mainly generated into journal for others to take advantages and improve as upcoming studies. The method is use to achieve the objective of the project that will accomplish a perfect result. In order to evaluate this project, the methodology based on System Development Life Cycle (SDLC), generally three major step, which is planning, implementing and analysis.

Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

To identify all the information and requirement such as hardware and software, planning must be done in the proper manner. The planning phase has two main elements namely data collection and the requirements of hardware and software.

Data collection:
Machine learning needs two things to work, data (lots of it) and models. When acquiring the data, be sure to have enough features (aspect of data that can help for a prediction, like the surface of the house to predict its price) populated to train correctly your learning model. In general, the more data you have the better so make to come with enough rows.

The primary data collected from the online sources remains in the raw form of statements, digits and qualitative terms. The raw data contains error, omissions and inconsistencies. It requires corrections after careful scrutinizing the completed questionnaires. The following steps are involved in the processing of primary data. A huge volume of raw data collected through field survey needs to be grouped for similar details of individual responses.

Data Pre-processing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

Therefore, certain steps are executed to convert the data into a small clean data set. This technique is performed before the execution of Iterative Analysis. The set of steps is known as Data Pre-processing. It includes –

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction

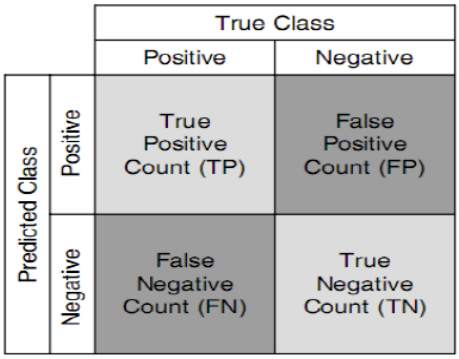
Data Pre-processing is necessary because of the presence of unformatted real-world data. Mostly real-world data is composed of –

- **Inaccurate data (missing data) -** There are many reasons for missing data such as data is not continuously collected, a mistake in data entry, technical problems with biometrics and much more.

- **The presence of noisy data (erroneous data and outliers) -** The reasons for the existence of noisy data could be a technological problem of gadget that gathers data, a human mistake during data entry and much more.

- **Inconsistent data -** The presence of inconsistencies are due to the reasons such that existence of duplication within data, human data entry, containing mistakes in codes or names, i.e., violation of data constraints and much more.

In this final phase, we will test our classification model on our prepared dataset and also predict the air quality on our dataset. To evaluate the performance of our created classification and make it comparable to current approaches, we use accuracy to measure the effectiveness of classifiers. After model building, knowing the power of model prediction on a new instance, is very important issue. Once a predictive model isdeveloped using the historical data, one would be curious as to how the model will perform on the data that it has not seen during the model building process. One might even try multiple model types for the same prediction problem, and then, would like to know which model is the one to use for the real-world decision making situation, simply by comparing them on their prediction performance (e.g., accuracy). To measure the performance of a predictor, there are commonly used performance metrics, such as accuracy, recall etc. First, the most commonly used performance metrics will be described, and then some famous estimation methodologies are explained and compared to each other. "Performance Metrics for Predictive Modeling In classification problems, the primary source of performance measurements is a coincidence matrix (classification matrix or a contingency table)". Above figure shows a coincidence matrix for a two-class classification problem. The equations of the most commonly used metrics that can be calculated from the coincidence matrix are also given in Fig. 3.1.



$$\text{True Positive Rate} = \frac{TP}{TP + FN}$$

$$\text{True Negative Rate} = \frac{TN}{TN + FP}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Figure 3.1: confusion matrix and formulae

As being seen in above figure, the numbers along the diagonal from upper-left to lower-right represent the correct decisions made, and the numbers outside this diagonal represent the errors. "The true positive rate (also called hit rate or recall) of a classifier is estimated by dividing the correctly classified positives (the true positive count) by the total positive count. The false positive rate (also called a false alarm rate) of the classifier is estimated by dividing the incorrectly classified negatives (the false negative count) by the total negatives. The overall accuracy of a classifier is estimated by dividing the total correctly classified positives and negatives by the total number of samples.

IV. RESULTS AND DISCUSSION

In this final phase, we will test our classification model on our prepared dataset and also predict the air quality on our dataset. To evaluate the performance of our created classification and make it comparable to current approaches, we use accuracy to measure the effectiveness of classifiers. After model building, knowing the power of model prediction on a new instance, is very important issue. Once a predictive model is developed using the historical data, one would be curious as to how the model will perform on the data that it has not seen during the model building process. One might even try multiple model types for the same prediction problem, and then, would like to know which model is the one to use for the real-world decision making situation, simply by comparing them on their prediction performance (e.g., accuracy). To measure the performance of a predictor, there are commonly used performance metrics, such as accuracy, recall etc.

V. CONCLUSION AND FUTURE SCOPE

Based on the data and results, the proposed methodology ofcharacterization of the air quality index using machinelearning-based predictive models is implemented successfully.A prototype composed of array of sensors are developed. Machine learning models are established with neural networkbeing the best, with an accuracy of 99.56% and a 0.0543logloss performance.

REFERENCES

[1] R. Tibshirani, "Regression shrinkage and selection via the lasso,"Journal of the Royal Statistical Society. Series B(Methodological),vol. 58, no. 1, pp. 267–288, 1996.

[2] M. Yuan and Y. Lin, "Model selection and estimation in regressionwith grouped variables," Journal of the Royal Statistical Society:Series B (Statistical Methodology), vol. 68, pp. 49–67, 2006.

[3] L. Li, X. Zhang, J. Holt, J. Tian, and R. Piltner, "Spatiotemporal interpolation methods for air pollution exposure," in Symposiumon Abstraction, Reformulation, and Approximation, 2011.

[4] Y. Zheng, F. Liu, and H.-P. Hsieh, "U-air: When urban air quality inference meets big data," in Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '13, 2013, pp. 1436–1444.

[5] World Health Organization (WHO), "7 million premature deaths annually linked to air pollution," Mar. 2014. [Online]. Available: http://www.who.int/mediacentre/news/releases/2014/airpollution/en

[6] Y.C. Wang and G.W. Chen, "Efficient Data Gathering and Estimationfor Metropolitan Air Quality Monitoring by Using Vehicular Sensor Networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 8, pp. 7234–7248,2017.

[7] Y. Li and J. He, "Design of an intelligent indoor air quality monitoringand purification device," in *2017 IEEE 3rd Information Technology andMechatronics Engineering Conference (ITOEC)*, 2017, pp. 1147–1150.

[8] G. O. Avendanoet al., "Microcontroller and app-based air quality monitoring system for particulate matter 2.5 (PM2.5) and particulate matter 1 (PM1)," in 2017 IEEE 9th International Conference onHumanoid, Nanotechnology, Information Technology, Communicationand Control, Environment and Management (HNICEM), 2017, vol. 5, pp. 1–4.

[9] J. Molka-Danielsen, P. Engelseth, V. Olesnanikova, P. Sarafin, and R. Zalman, "Big Data Analytics for Air Quality Monitoring at a Logistics Shipping Base via Autonomous Wireless Sensor NetworkTechnologies," 2017 5th Int. Conf. Enterp. Syst., pp. 38–45, 2017.

[10] Y. Wu et al., "Mobile Microscopy and Machine Learning Provide Accurate and High-throughput Monitoring of Air Quality," in Conference on Lasers and Electro-Optics, 2017.

Authors Profile

Mr. ASHOK SHAH pursuing Bachelor of Computer Science and Engineering from Visvesvaraya Technological University, Belagavi.

*Mr. PRAYASH RIMAL*pursuing Bachelor of Computer Science and Engineering from Visvesvaraya Technological University, Belagavi.

Mr. DEV BHASKER SINGH pursuing Bachelor of Computer Science and Engineering from Visvesvaraya Technological University, Belagavi.

*Mr. JAGADEESH B N*pursued Bachelor of Computer Science and Engineering from Visvesvaraya Technological University, Belagavi in 2006 and Masters of Technology from Visvesvaraya Technological University in 2009 and Currently working as Assistant Professor in Dept. of Computer Science