Labelling Data for Correlation Pollution Dataset by Using Machine Learning

Huda W Ahmed¹, Dr Jameelah H Alamire²

¹ Institute of informatics for Postgraduate studies, Iraq.
² Computer Science Department, College of Science, Mustansiriyah University, Iraq.
*Email: huda_wadah@yahoo.com; dr.jameelahharbi@gmail.com

Abstract. The significant problem in the environment is the pollution of air which affects all the areas in our world. The solution of this problem is excluding by using Internet of Things (IoT) technology which designed system that monitors the parameters that affect the pollution in the environment. The handling and operation of large quantities of information from a wide range of sources is challenges in IoT technology. The purpose of this study is to decrease the dimensions in IoT systems so we suggest two models for IoT system and also aims to build rules for Air quality classification by defining pollution levels. In this study, we suggested machine learning techniques to predict an air pollution levels based on a dataset consisting of daily weather, also we simplified our classification models to solve this problem by using binary classification which is to classify the air pollution level into "High pollution" and "Simple pollution" class labels. The decision tree J48 Algorithm gives the best outcomes in the first model than the second models.

1. Introduction

Air pollution can be described as the entry into the earth's environment of dangerous substances such as toxic chemicals. These pollutants influence the atmosphere and living organisms directly or indirectly. The environmental effects of air pollution range from acid rain, global warming, ozone depletion and climate alter leading to drought and vegetation harm [1]. Global warming and climate change has become the biggest threat now facing the whole universe. Due to multiple gasses such as carbon dioxide, sulfur dioxide, Carbon monoxide, Nitrous oxide, Methane, Ozone, etc., the contribution to the rising temperature of the earth is both dangerous to all biotic components at ground level and their contribution to world pollution is enormous [2]. IoT's key is connectivity of physical objects via the Internet via wireless or wired technology. IoT offers a broad variety of alternatives such as tracking the environment, intelligent transport system, e-health, smart grid, smart homes, etc.[3& 4]. IoT technology is used to track the quality of environmental parameters such as air, noise, temperature, humidity, light and various gases. Monitoring pollution concentrations are used in manufacturing environments or in specific areas of interest. It enables officials take action when the pollution crossing the defined level [5]. The word IoT refers to a scheme that includes linked devices that collect data, connect to the Internet or local networks, produce analytics, and adapt reactions based on the network's data analysis. Connected devices capturing or generating huge amounts of data every day, together with the applications and services used to interpret, analyze, predict and take action based on the information received [6].
This study is structured as follows: Discussed the heading ‘Literature Review’ in Section 2. Section 3 explained under the heading ‘Background theory’. Section 4 explained the proposed method. Section 5 explained implementation of suggested models. Section 6 presents the observed results and Section 7 the conclusion of this study.

2. Literature Review
In the research [7] on the Mohammed al-Qasim highway adjacent to the University of Technology was performed in Baghdad. Results of the research showed that some kinds of sulfur compounds (SO2) were high at hazardous health-threatening levels due to the elevated sulfur content in Iraqi petrol and also elevated nitrogen dioxide (NO2) and carbon monoxide (CO) concentrations can make the area susceptible to smog formation and harm human health.

In the paper [1] the authors detect the presence of harmful gases by designing a new device. If the amount of polluting gases exceeds a certain predefined limit the new device give an indication. The new device can be executed in any industry, and can control the environment and keeps the atmospheric balance.

In article [5] the authors designed user friendly monitoring system which appeared simulation output for the carbon dioxide gas, noise pollution, relative humidity and temperature in the Environment. The system contains senses devices, Arduino board and Wi-Fi module to monitor the environment parameters. The results of the sensors migrated to the cloud and can be appeared through internet.

In the document [8] focuses on air assessment based on accessible information from multiple air pollutants such as NO2, SO2, CO and O3 with corresponding values for the Air Quality Index (AQI). This paper uses Decision tree J48 algorithm and Naïve Bayes to predict the health concern. Air Quality Index (AQI) categories are good, moderate, (unhealthy for sensitive groups), unhealthy, very unhealthy. The result showed that the Decision tree algorithm gives an accuracy of 91.9978 % which is more than Naïve Bayes algorithm viz. 86.663%.

In the paper [9] the aim is to find the way to monitor the air quality through the use of integrated gas sensors using DHT 11 temperature and relative humidity sensor, MQ2, MQ5 and MQ135 gas sensors and building five models using Machine learning algorithms. This models are k-nearest neighbours (KNN), support vector machine (SVM), Naïve-Bayesian classifier, random forest and neural network. The results show that the neural network obtains the best accuracy 99.56% than the others models.

3. Background Theory
In this section we describe the background theory about this study that involves:

3.1. IOT
The advanced and omnipresent IoT applications drive a big quantity of data to grow. Data produced by heterogeneous machines is difficult to evaluate and data can consist of various aspects. Dimension reduction is one of the main techniques of reducing a specified dataset's amount of variables. As data comes from heterogeneous devices, the amount of features can differ for IoT systems. New algorithms must be intended to address the dynamic nature of IoT systems [4].

3.2. Machine Learning
Machine learning provides many approaches that can offer new opportunities for prediction of air pollution. First building machine learning models then training existing data sets on these models [10]. In this study we apply some Machine Learning techniques to predict air pollution levels based on a dataset.

3.2.1 Dimensionality Reduction
Dimension reduction relates to the method of transforming large-scale data into smaller-scale data, ensuring that it concisely conveys comparable information to achieve better features for classification, clustering or regression characteristics [4]. IoT devices, reading, writing and processing a big amount of data are challenging. Therefore, effective and robust techniques are needed to handle a big amount
of data on IoT devices. In addition, heterogeneous data produced by distinct kinds of IoT devices need to be integrated. Also it is necessary to extract the main information from a big quantity of data [11]. Principal Component Analysis (PCA) is a method for dimensionality reduction to reducing the larger number of variables in to the smaller number of variables called principal components in a new set [12]. PCA is a traditional technology for statistical analysis and can be used not just to reduce data dimension, avoid redundant data in data and correlation between variables, and reduce computational complexity, but also to provide significant data such as data mean level, the biggest data average direction. PCA can also be utilized for extraction of features [13]. There are several benefits for using dimension reduction in IoT scheme such as: Dimension decrease can be useful in reducing computing energy for data analytics as well as necessary storage space. Also the communication costs can be decreased by using IoT scheme; Less-sized data will cost less overhead communication and can be implemented current algorithms to less-sized data. With fewer dimensional the data visualization is much easier and possible to observe the data patterns more obviously and possible noise reduction [4].

3.2.2 Classification Method
Classification is the search for a model that defines a class label so that it can be used to predict an unknown class label [14] . The classification methods are called monitored or supervised learning. Which is often offers significant application-specific value, as class labels may represent significant interesting characteristics. Typically, classification algorithms comprise two stages:
• Training stage: a model is built from the training instances in this stage (building model).
• Testing Stage: The model is used to assign an unlabelled test instance with a label and predict an unknown class labels [14 &15]. The Classification method applied in this study is Decision Tree method and building its model.

3.2.3 Decision Tree
A decision tree is a tree-like arrangement with a node of root, branches, and nodes of leaf. Each internal node defines an attribute test condition, each branch indicates the result of the condition of the internal node, and each leaf node displays a class label. For classification, the decision tree is used. Decision tree is then built to test set for class label assignment [8].

4. Proposed Method
This section includes the proposed method to deal with the dataset:

4.1. Dataset Description
The experiment is applied on the real Iraqi air pollution data set for the year 2012 for three stations (Alwazyrea, Alandulas, Algadrea). These station are located at the city of Baghdad. The data contains the various parameters like concentration of CO, NO2, SO2 and the Meteorological data consisting of temperature and humidity, this data used to covers the per hour of various parameters that are defined in the below Table 1.
4.2. Dataset Pre-processing
The data are collected and classify according to Iraqi Air Quality standards in Table 2, which is defined by the Ministry of the Planning in Iraq.

| Parameters | Values            |
|------------|-------------------|
| Concentrations |                  |
| CO         | 0.01 ,.., 0.56   |
| SO2        | 0.0 ,.., 0.198   |
| NO2        | 0.0 ,.., 0.19    |
| Meteorological |            |
| Temperature| 12 ,.., 47       |
| Humidity   | 25,.., 96        |
| Class labels |            |
| Simple     | 0                |
| High       | 1                |

Table 2. Iraqi Air Quality Standards

| Pollutant               | Ambient Air Standard (PMM) |
|-------------------------|----------------------------|
| Carbon Monoxide (CO)    | 0.26                       |
| Sulfur Dioxide (SO2)    | 0.14                       |
| Nitrogen Dioxide (NO2)  | 0.05                       |

The concentrations of gases which located in Table 2. Defined as information in [16 & 17 & 18 & 19].

We found that all the datasets are polluted, which means the concentration of one gas or two are crossed the safe limit or the threshold that is placed by Iraqi Air Quality Standards.

So we suggest classify data according to this way into two classes:
Class 0 for simple or acceptable pollution (one gas crossed the safe limit).
Class 1 for high or unacceptable pollution (two gases crossed the safe limit).

As shown in the below in the Table 3.
Table 3. Rules for classify Pollution level

| Rule # | Rules                                                                 | Pollution level | Class label |
|-------|-----------------------------------------------------------------------|----------------|-------------|
| 1     | IF CO > 0.26 & SO2 < 0.14 & NO2 < 0.05                                 | Simpl          | 0           |
| 2     | IF CO < 0.26 & SO2 > 0.14 & NO2 < 0.05                                 | Simpl          | 0           |
| 3     | IF CO < 0.26 & SO2 < 0.14 & NO2 > 0.05                                 | Simpl          | 1           |
| 4     | IF CO > 0.26 & SO2 < 0.14 & NO2 > 0.05                                 | Simpl          | 1           |
| 5     | IF CO > 0.26 & SO2 > 0.14 & NO2 < 0.05                                 | Simpl          | 0           |
| 6     | IF CO < 0.26 & SO2 > 0.14 & NO2 > 0.05                                 | Simpl          | 0           |

The total records in the dataset are 4544 and the information from the output was marked as one or zero. One relates to increased levels of pollution, whereas zero relates to low levels of pollution. So the complete amount marked as zero is 3372, while the other 1172 points are marked as one as show in figure 1.

![Figure 1. Levels of pollution in class labels](image)

5. Suggested Models

We suggest applied system for IOT with using two models and then compared between the results of them:

- The first model contains one phase which performed the classification algorithm (Decision Tree) directly to the real Iraqi air pollution dataset to predict the level of pollution. As shown in figure 2.
The second model contains two phases:

In the first phase implement PCA algorithm due to importance of the time factor to the IoT technology with one output feature, three output features and five output features, and in the second phase performed the classification algorithm (Decision Tree) on the real Iraqi air pollution dataset to classify and predicate the pollution level. As shown in figure 3.

![Figure 2. First model](image1)

![Figure 3. Second model](image2)

6. Results

In this section we discuss the results of the study:

6.1. Statistic Correlation Results
The study showed that there is a correlation between the concentration of CO, NO2, SO2 and Temperature and Humidity as show in the Table 4.

| Table 4. Statistic correlation results |
As shown in the Table 4, there is a negative correlation between CO and temperature (means CO increased when the temperature decrease) and positive correlation with humidity (CO increased when humidity increased).

Also there is a negative correlation between SO2 and temperature and negative correlation with humidity. And finally there is a positive correlation between NO2 and temperature and negative correlation with humidity. In the figure 4. The Correlation between gases and Temperature and Humidity.

### Table 4. Concentration Temperature Humidity

|          | Temperature | Humidity |
|----------|-------------|----------|
| CO       | -0.2343     | 0.2723   |
| NO2      | 0.1994      | -0.0479  |
| SO2      | -0.0573     | -0.0191  |

6.2. Evaluation Models

The parameters that considered when evaluating the two models by using the data mining algorithm (Decision Tree) are Accuracy, Kappa, Mean Absolute Error and finally Root Mean Squared Error.  
Table 5. Shows the results related to the evaluating of the two models.

| Parameters          | Model 1 | Model 2 |
|---------------------|---------|---------|
|                     | One output feature | Three output features | Five output features |
| Accuracy            | 99.92%  | 81.07%  | 96.184%  | 94.057%  |
| Kappa Statistic     | 0.9981  | 0.4588  | 0.9006   | 0.8448   |
| Mean Absolute Error | 0.0007  | 0.2922  | 0.0447   | 0.0684   |
| Root Mean Squared Error | 0.0271 | 0.3793 | 0.1895 | 0.2362 |
7. Conclusions
In this study we suggest system for IoT with using two models for classifying and predicating the data that’s comes from IoT sensors. These sensors give real-time data to predicate the level of the pollution. The data that are used are local and real data for Iraqi Air pollution to year 2012. The dataset gathers from three stations (Alwazyrea, Alandulas, and Algadorea) in Iraq and then aggregated. The labels are placed according to the standard concentrations in Table1.

In the first model executed without using PCA algorithm which directly implemented the classification algorithm, in the second model using PCA algorithm due to the time factor is important for IoT with one output feature, three output features, five output features and then implemented the classification algorithm (Decision Tree). The results of the study are showed:

- The first model gives a better understanding of the pollution levels to the public and higher accuracy than the second model.
- The study showed that there is a correlation between the concentration of CO, NO2, SO2 and the temperature and humidity.
- The atmosphere of Iraq is polluted, which found that the concentration of one gas or two are crossed the safe limit or threshold that is placed by Iraqi air quality standards, therefore we suggest classify the data into two classes: simple pollution and high pollution.
- This study about by using of PCA algorithm didn’t give its benefits or advantages which have low accuracy due to the minimum number of the features in the dataset. While we recommended using PCA algorithm with IOT when the dataset have big number of features.

On the other hand, the data set in this paper is no large enough. Air quality is a long-term formed problem and it is better to use a large data covering a variety of years and locations. Furthermore, beside the meteorological and traffic factors.

References:
[1] Sidharth A, Harivishnu B, Vinesh T K, Sooraj S and Geena P 2017 Automated gas pollution detection system 2nd International Conference for Convergence in Technology (I2CT)483-486
[2] Murtala G Murgan, Abdul G Hamid 2016 Towards Regulation of Devastating a Viation Emission: An Analysis of the Contaminants of Aviation Emission, its Global Environmental Impact and the Legal Regime J. Asian African Soc. Sci. Humanit 2 55-73.
[3] Waleed E, Mohamed I 2015 Machine-to-machine communications in cognitive cellular systems IEEE International Conference on Ubiquitous Wireless Broadband (ICUWB) 1-5.
[4] Waleed E and Alagan A 2019 Internet of Things for Smart Cities: Technologies, Big Data and Security (Switzerland:Springer)
[5] A N Chaudhari, G A Kulkarni 2017 IOT based environmental pollution monitoring system International Research Journal of Engineering and Technology (IRJET) 4 1823. staff of the World Bank Group 2017 Internet of Things, The New Government to Business Platform A review of opportunit.
[6] Miqdam T Chaichan, Hussien A Kazem and Talib A Abed 2016 Traffic and outdoor air pollution levels near highways in Baghdad, Iraq Springer 20
[7] Ranjana W Gore, Deepa S. Deshpande 2017 An approach for classification of health risks based on air quality levels IEEE 58-61.
[8] Timothy M Amado, Jennifer C Dela Cruz 2018 Development of Machine Learning-based Predictive Models for Air Quality Monitoring and Characterization IEEE Region 10 Conference 668-672.
[9] Ziyue G, Richard O Sinnott 2018 Prediction of Air Pollution through Machine Learning Approaches on the Cloud. IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT) 51-60.
[10] Marco S 2016 The internet of things: Opportunities and challenges for distributed data analysis SIGKDD Explorations 18 15–34
[12] R. Indhumathi, Dr. S. Sathiyabama 2010 Reducing and clustering high dimensional data through principal component analysis International Journal of Computer Applications 11
[13] Lina Y, Huajun C, Bo T and Ging G 2018 Data mining-based indoor air quality monitoring system in the era of big data IOP Conference Series: Earth and Environmental Science 208
[14] Obuandike Georgina N, Audu Isah and John Alhasan 2015 Analytical Study of Some Selected Classification Algorithms in WEKA Using Real Crime Data (IJARAI) International Journal of Advanced Research in Artificial Intelligence 4
[15] Charu C. Aggarwal 2014 Data classification: algorithms and applications CRC press (New York, USA)
[16] Areaj kh alrawi, Rana H 2013 Impact environment of electric generation in Baghdad city-alkarada district 903 J. Educ. Coll. Wasit University 14 294-314.
[17] Shno M Ali 2019 Study the environmental effects of private electrical generators for some locations in Kirkuk city Sulaimani Journal for Engineering Sciences 6.
[18] Salih N Majid 2011 Valuation of Ambient Air Pollution: A Study of some Urban Areas in Sulaimania City and its Surrounding / Kurdistan Region of Iraq Iraqi Kurdistan Region Ministry of Higher Education and Scientific Research of Sulaimania.
[19] Ministry of Planning- Central Organization for Standardization and Quality Control 2010 Iraqi Specification no 4059 for Iraq.