International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

Analysis of Fine-Grained Air Quality and Prediction of Air Pollution

Shashi Rekha G [1], Likhita R [2], Arpitha B Y [3], Ananya R [4], Aishwarya S [5]

[1]*Assistant Professor, Dept. of Computer Science Engineering, Sapthagiri College of Engineering, Bangalore, India.*
[2,3,4,5]*Final Year Students, Dept. of Computer Science Eng., Sapthagiri College of Eng., Bangalore, India.*

Abstract - *Air pollution is one of the biggest threats for the environment and affects everyone: humans, animals, crops, cities, forests and aquatic ecosystems. Air pollution is usually caused due to more usage of electricity, fuel and transportation. The solutions to these topics can provide extremely useful information to support air pollution control .The main objective of this system is to use propose general and effective approach to predict the air quality and give suggestion which will create the impact on society. The interpolation, prediction, and feature analysis of fine-gained air quality are three important topics in the area of urban air computing. Most of the existing work solves the three problems separately by different models. Since there are insufficient air-quality-monitor stations in a city due to the high cost of building and maintaining such a station, it is expensive to obtain label training samples when dealing with fine-gained air quality. The dataset is collected from central pollution control board. The proposed approach detects the various stages of air pollution such as good, moderate-demented and non-demented using Random forest algorithm and SVM algorithm.*

Key Words: Interpolation, SVM, random forest.

1. INTRODUCTION

Air pollution is one of the biggest threats for the environment and affects everyone: humans, animals, crops, cities, forests and aquatic ecosystems. Air pollution is usually caused due to more usage of electricity, fuel and transportation. Air pollution has become an alarming environmental issue globally due to rapid urbanization and industrialization. The main atmospheric pollutants in gas composition contains Sulphur dioxide, Nitrogen oxide, Nitrogen Dioxide, Ozone, Carbon monoxide, Carbon dioxide and so on. Among different air pollutants, airborne particulate matter (PM) with diameters less than 2.5micrometers (PM2.5) has significant harmful effects on the human body as these particles are capable of transmitting hazardous chemicals into the human lung and blood and cause cardiovascular, respiratory and cerebrovascular diseases, reduced lung functions, and heart attacks. Recent studies have shown substantial evidence that exposure to atmospheric pollutants has strong links to adverse diseases including asthma and lung inflammation. In last decade there is a lot of improvement in the techniques of air quality monitoring and forecasting. Currently, air quality monitoring methods are mainly based on monitoring stations, which are not available to the majority of regions because of the high setup cost and expensive sophisticated sensors. The interpolation, prediction, and feature analysis of fine-gained air quality are three important topics in the area of urban air computing. A good interpolation solves the problem that there are limited air-quality-monitor-stations whose distribution is uneven in a city. A precise prediction provides valuable information to protect humans from being damaged by air pollution. A reasonable feature analysis reveals the main relevant factors to the variation of air quality. In general, the solutions to these topics can extract extremely useful information to support air pollution control, and consequently generate great societal and technical impacts. However, there exist several challenges as the related data have some special characteristics. First, since there are insufficient air-quality-monitor stations in a city due to the high cost of building and maintaining such a station, it is expensive to obtain label training samples when dealing with fine-gained air quality. Second, the labelled data of the air-quality-monitor-stations are incomplete, and there exist lots of missing labels of the historical data in some time periods for some stations. The reason for the incomplete labels is related to the air quality monitor devices. In general, each station only has one monitor device which needs to be maintained at intervals, thus there will be no outputs for the station when the device is being maintained, or has other problems. Third, the kinds of urban air related data are various for the development of data acquisition technologies. In the era of information, the use of intelligent algorithms to deal with data in the mainstream direction. The data of air quality comes from the sensor and the amount of data is too large. In recent years, some intelligent algorithms in the air quality assessment are used, such as artificial neural network pattern recognition is used successfully neural network is applied to the evaluation system, but the convergence speed of BP neural networks is slow. The disadvantage of this is that it can be easy fall into local optimization. So to overcome these, the algorithms like random forest algorithm and Support vector machine (SVM) can deal with sorting of large amount of data.

Fig -1: The image shows continuous air quality changing of 16 days at same location.

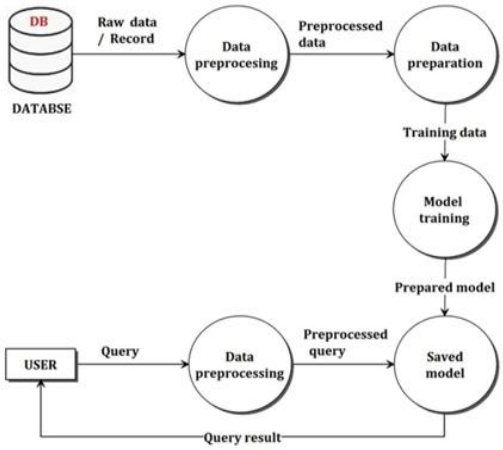

Fig -2: System architecture

2. METHODOLOGY

Prediction of air pollution will help environment and nature from being destroyed. The detection of air pollution is done by wireless sensor networks, which are used as active research area. The detection by these techniques is time consuming. Hence we are applying the machine learning technology to predict the Air pollution. The proposed approach detects the various stages of Air pollution using Random forest algorithm Support vector machine (SVM).Both the algorithm are used for more accuracy rate. It reduces the time required to predict the output and can be used for real time predictions. We describe the datasets used in this study and how the data was pre-processed before the machine learning task. Feature extraction using principal component analysis and feature selection techniques were also employed. After the data preprocessing is done, the efficient machine learning algorithms that are random forest and SVM are applied to predict the pollution rate and categorized it into different levels.

2.1 Data preprocessing

Data Preprocessing is a technique that is used to convert the raw data into a clean data set. The data is gathered from different sources it is collected in raw format which is not feasible for the analysis. Data preprocessing consists of various steps like: Data Cleaning, Data Integration, Data Transformation, and Data Reduction. Data Preprocessing is necessary because of the presence of unformatted real-world data. Mostly real-world data is composed of inaccurate data, the presence of noisy data and Inconsistent data. The preprocessing is also important to speed up training such as clustering and scaling technique. Data preprocessing is a proven method of resolving such issues.

2.2 Feature selection and extraction

Feature selection is the process of selecting a subset of relevant features for use in model construction. Here relevant features would be the only those parameters which causes air pollution. It is an effective strategy to optimize the predictive performance of machine learning algorithms.

2.3 Random Forest Algorithm

Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because its simplicity and the fact that it can be used for both classification and regression tasks. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

Random Forest is a supervised learning algorithm; it creates a forest and makes it somehow random. The “forest” it builds, is an ensemble of Decision Trees, most of

the time trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result.

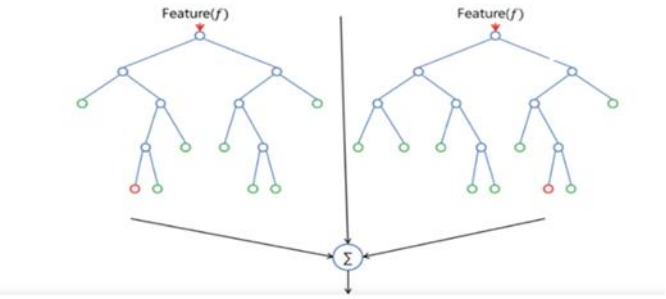

Fig -3: Random forest

One of the great qualities of the random forest algorithm is that it is very easy to measure the relative importance of each feature on the prediction. Hyper-parameters in random forest are either used to increase the predictive power of the model or to make the model faster. The number of hyper-parameters is also not that high and they are straightforward to understand. The Random forest algorithm is considered as a very handy and easy to use algorithm.

2.4 Support Vector Machines (SVM)

A Support Vector Machine (SVM) is a supervised machine learning algorithm that can be employed for both classification and regression purposes. SVMs are based on the idea of finding a hyper plane that best divides a dataset into two classes.

The first selected mining technique is SVM that can be used as linear and nonlinear. "In machine learning, SVMs is supervised learning model with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier.

Support vectors are the data points nearest to the hyper-plane, the points of a data set that, if removed, would alter the position of the dividing hyper-plane. Because of this, they can be considered the critical elements of a data set. The distance between the hyper-plane and the nearest data point from either set is known as the margin. The goal is to choose a hyper plane with the greatest possible margin between the hyper-plane and any point within the training set, giving a greater chance of new data being classified correctly. The data will continue to be mapped into higher and higher dimensions until a hyper-plane can be formed to segregate it. The main advantage is their accuracy of 89%, when both the algorithms are used.

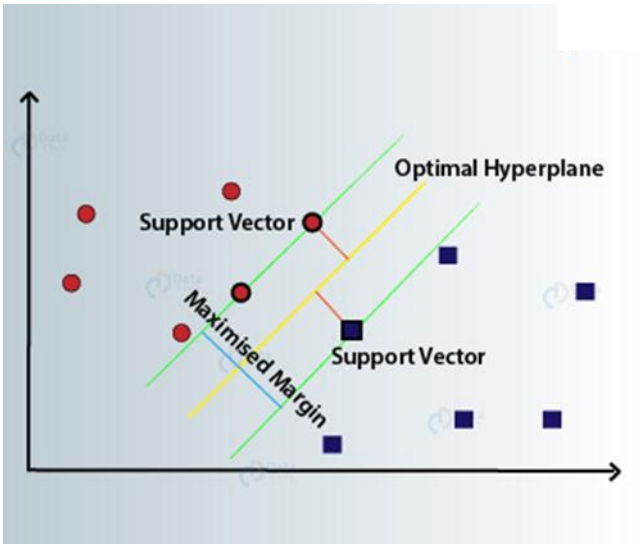

Fig -4: Support Vector Machines

3. CONCLUSIONS

In this paper, we used machine learning algorithms along with data preprocessing and principal component analysis and feature selection techniques to classify required parameters from whole dataset using data collected from Central Pollution Control Board (CPCB).

The SVM and random forest algorithm has been employed to carry out the prediction. Compared with traditional methods, the proposed method has achieved about 30% of improvement on the classification accuracy of 89%. The best applications of this work is providing crucial information to support Air Pollution Control and in generation of great societal aware and technical impact due to air pollution.

REFERENCES

[1] K. B. Shaban, S. Member, A. Kadri, and E. Rezk, "Urban Air Pollution Monitoring System with Forecasting Models," IEEE SJ, vol. 16, no. c, pp.1–9, April 2016.

[2] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," Journal of the Royal Statistical Society:SeriesB(Statistical Methodology), vol. 68, pp. 49–67,2012.

[3] K.P .Singh, S.Gupta, and P.Rai, "Identifying pollution sources and predicting urban air quality using ensemble learning methods," Atmospheric Environment, vol. 80, pp. 426 – 437, 2013.

[4] C. Zhang and D. Yuan, "Fast fine-grained air quality index level prediction using random forest algorithm on cluster computing of spark," Proc. -2015 IEEE 12th Int. Conf. Ubiquitous Intell.Comput. f20, pp. 929–934, 2016.