Research on Urban Air Quality Prediction Based on Ensemble Learning of XGBoost

Haotian Jing[1], Yingchun Wang[2]

[1]School of management, Tianjin University of Technology, Tianjin 300384, China, 245952376@qq.com
[2]School of management, Tianjin University of Technology, Tianjin 300384, China, ycwang@sina.com

Abstract. In recent years, with the rapid development of China's economy and the continuous improvement of people's quality of life, air pollution caused by a large amount of energy consumption has become increasingly serious. Air quality index (AQI) has become an important basis to measure air quality. At present, the research on air quality assessment and prediction methods has become increasingly active at home and abroad, which is of great significance to guide people's production and life. In this paper, Taking Shijiazhuang, Hebei Province as an example and using the XGBoost model of the machine learning ensemble algorithm, regression fitting was performed on the six pollutant concentrations that currently mainly affect air quality, and the hourly prediction of AQI was achieved.The trained model has lower mean absolute error (MAE) and higher correlation coefficient (R-square), which improves the prediction ability of urban air quality prediction, provides a new idea for urban air quality prediction, and has a broad application prospect in the future urban air quality prediction.

1 Introduction

In recent years, China's economy has developed steadily and rapidly, the industrialization of urban and rural areas and the living standards of residents have been greatly improved, and the scale of industry and transportation has continued to expand. However, the problem of disharmony between environment and development has become increasingly prominent. The consumption of fossil fuels is increasing, and the emission of automobile exhaust, industrial emission, construction dust, waste incineration and other exhaust gases is increasing. As a result, the air pollution problems caused by the main pollutants such as inhalable particles (PM10, PM2.5), sulfur dioxide (SO2), nitrogen dioxide (NO2), nitric oxide (no), nitrogen oxide (no), carbon monoxide (CO), ozone (O3) are becoming increasingly serious, It has gradually become a major livelihood issue that the people are increasingly concerned about[1]. In recent years, many northern cities frequently have different degrees of haze weather, which not only has a negative impact on people's normal production, life, work and learning, but also hinders the sustainable development of society. Shijiazhuang, the capital of Hebei Province, as a key pollution city in China, how to make correct prediction and assessment of future air quality changes, and finally achieve effective control of regional ambient air quality has become an important research topic of practical significance.

On the other hand, with the continuous improvement of the urban ambient air quality monitoring system, the monitoring data is growing rapidly, and a large amount of historical monitoring data has been accumulated, which provides an important basis for the analysis and control of the urban ambient air quality. However, the traditional data processing methods and means have not been able to effectively analyze and process the massive and high-dimensional air quality monitoring data. New technical support is urgently needed to realize the analysis and utilization of the existing data. In view of the above background, this paper studies XGBoost model based on machine learning boosting ensemble algorithm, analyzes and cleans the historical monitoring data of air quality, obtains the prediction model through machine training, so as to grasp the trend of air quality change, and provides scientific and reasonable decision-making information.

2 The current research

At present, the prediction of air quality at home and abroad can be roughly divided into statistical model-based method, machine learning based method and deep learning method.

Based on the prediction method of statistical model, according to the principle of statistics, analyze the concentration of pollutants and meteorological parameters or other pollutants, find out the internal relationship between them, and establish the pollutant concentration prediction model according to the relationship[2]. For example, Ross. Z. et al. Developed regression equation to predict PM2.5 concentration of air monitoring points in New York City by using traffic and land-use model data[3]; Li et al. Used univariate linear

regression model to draw a conclusion that PM2.5 concentration has significant negative correlation with different meteorological parameters (temperature, rainfall, wind speed)[4]; Jiao et al. Took various monitoring indicators as prediction variables, based on multivariate analysis Regression analysis model predicts AQI and analyzes the influencing factors of air quality[5].

With the research and application of machine learning and neural network, air quality prediction based on machine learning and deep learning has become an important research field. Yeganeh. B. et al. Studied the support vector machine prediction model based on CO concentration measurement and partial least square method as data selection tool for air pollution prediction[6]; Zhang Xilai et al. Proposed a dynamic adjustment model based on single time series data for PM2.5 concentration prediction, effectively combined the exponential smoothing method and Markov model, thus improving the prediction accuracy Accuracy[7]. Zong Xiaoping et al. Proposed a prediction model of optimizing parameters of support vector regression machine by genetic algorithm to predict PM2.5 concentration[8].

3 Data and model

3.1 Data source

The air quality data of Shijiazhuang used in this paper are all from the national real-time urban air quality release platform of China Environmental Monitoring Center. The monitoring period is from January 2, 2015 to December 31, 2019, and the data sampling interval is one hour. Through data cleaning and preprocessing, 33708 pieces of effective data are obtained, as shown in the Figure 1.

datatime	AQI	PM2.5	PM10	SO2	NO2	O3	CO
2019/12/19 7:00	77	53	102	17	57	9	1.4
2019/12/19 8:00	85	59	109	16	55	10	1.62
2019/12/19 9:00	88	60	118	16	54	11	1.9
2019/12/19 10:00	88	61	119	16	50	13	1.82
2019/12/19 11:00	84	61	99	18	47	18	1.72

Fig.1. Part of Air Quality Concentration in Shijiazhuang on Dec 19, 2019

The original data used in this paper consists of two parts:(1) The model indexes are mainly six main pollutant concentrations in air quality data, including inhaled particulate matter (PM2.5, PM10), sulfur dioxide (SO2), nitrogen dioxide (NO2), ozone (O3), carbon monoxide (CO), in which the unit of CO is mg / m3, and the unit of other pollutants is μg / m3;(2) Model target values:air quality index (AQI).

3.2 XGBoost model

The air quality prediction based on machine learning method, because of its flexible non-linear modeling ability, is often superior to the traditional statistical method in the prediction effect, but a single machine learning model often depends on expert knowledge and feature engineering to improve the prediction effect of the model. Boosting ensemble learning is a new rising machine learning mode. Its basic idea is to constantly use a "weak" classifier to make up for the shortcomings of the previous "weak" classifier, and finally form a "strong" classifier serially. Its formula is as follows:

$$y_i^{(t)} = y_i^{(t-1)} + f_t(x_i) \qquad (1)$$

Among them, $y_i^{(t)}$ is the model prediction of round T, $y_i^{(t-1)}$ retains the model prediction of round t − 1, and $f_t(x_i)$ is the newly added function

XGBoost model is one of the boosting ensemble algorithms, which is based on the lifting tree model, so it ensembles many tree models together to form a strong classifier. At the same time, XGBoost model is improved on the basis of GradientBoostingDecisionTree (GBDT), making it more powerful and applicable to a wider range. Therefore, XGBoost model has the advantages of fast computing speed, strong model generalization ability, and significant model improvement effect. It is often used in some big data competitions, which can be used for both classification and regression problems.

4 Model training

4.1 K-fold cross validation

In order to ensure the randomness of the training data and the generalization ability of the model, in XGBoost model training, k-fold cross validation of the data set is carried out - the original data are divided into k groups, each subset data is taken as a validation set, the rest k-1 subset data is taken as a training set, K models are obtained through the training data, and then K models are used to fit their validation sets, The average value of the evaluation indexes of the K model is taken as the final performance index of the k-fold cross validation model.

In this paper, k = 10, i.e. the original air quality data set is divided into 10 parts, of which 9 parts are used as training data for model training in turn, and 1 part is used as test data for model evaluation

4.2 Model parameters selection

Compared with expert knowledge or feature engineering, parameter selection can make XGBoost model fit our data better and faster, and make the model have better generalization ability and prediction effect. In this paper, the model training is based on the open source Python packet manager Anaconda environment, using the XGBRegressor regression algorithm in the XGBoost toolkit. The tuning parameters include: max_depth=5, learning_rate=0.1,n_estimators=500,objective='reg:squarederror'

5 Model evaluation and result analysis

5.1 Evaluation index

In this paper, two commonly used indexes to evaluate the prediction performance of regression model are selected as follows:

(1) Mean absolute error (MAE): The mean absolute error (MAE) reflects the absolute difference between the predicted value and the actual value. The smaller the value is, the higher the prediction accuracy is.

(2) Square correlation coefficient(R^2): The correlation coefficient R2 is used to measure the correlation degree between the predicted value and the actual value. The larger the value is, the better the prediction effect of the model is.

5.2 Result analysis

The air quality data of Shijiazhuang from January 2, 2015 to December 31, 2019 are brought into XGBoost model under 10 Kold cross validation, and the model evaluation results are as follows:

```
mean absolute error is:  2.2095051822004352
R Squared is:  0.9951428345511243
```

Fig.2. Model Evaluation Results

Therefore, XGBoost ensemble algorithm has significant effect on prediction accuracy, error rate and interpretability. It can effectively predict the urban air quality index (AQI) through the current six air pollutant concentrations, so as to grasp the future trend of AQI, better guide people's production practice, and provides scientific and reasonable decision basis for the prevention of urban air pollution and the proposal of effective measures.

Reference

1. Xu xuran, Tu JUANJUAN. Air quality prediction system based on decision tree algorithm [J]. Electronic Design Engineering,2019,27(09):39-42.
2. Ren Cairong. Urban PM (2.5) concentration prediction based on parallel random forest [D]. Taiyuan University of Technology,2018.
3. Zev Ross, Michael Jerrett, Kazuhiko Ito, Barbara Tempalski, George D. Thurston. A land use regression for predicting fine particulate matter concentrations in the New York City region[J]. Atmospheric Environment,2006,41(11).
4. Atmosphere Research; Investigators at China Meteorological Administration Discuss Findings in Atmosphere Research (Variations in PM10, PM2.5 and PM1.0 in an Urban Area of the Sichuan Basin and Their Relation to Meteorological Factors) [J]. Science Letter,2015.
5. Zhang Xilai, Zhao Jianhui, Cai Bo. Dynamic adjustment prediction model for PM2.5 single time series data [J]. Journal of automation,2018,44(10):1790-1798.
6. Jiao Dong, sun Zhihua. Regression analysis of air quality index [J]. Journal of Ocean University of China (NATURAL SCIENCE EDITION),2018,48(S2):228-234.
7. B. Yeganeh, M. Shafie Pour Motlagh, Y. Rashidi, H. Kamalan. Prediction of CO concentrations based on a hybrid Partial Least Square and Support Vector Machine model[J]. Atmospheric Environment,2012,55.
8. Zong Xiaoping, Wu Zihan, Liu Yan. SVR haze prediction model based on genetic algorithm optimization [J]. Journal of Hebei University (NATURAL SCIENCE EDITION),2016,36(03):307-3