A Bagging-GBDT ensemble learning model for city air pollutant concentration prediction

Xinle Liu1,2,*, Wenan Tan1 and Shan Tang1,2
1 College of Engineering, Shanghai Polytechnic University, Shanghai, 201209, China
2 Shanghai Zhi Pan Intelligent Technology Co., Ltd., Shanghai, China
*Corresponding author's e-mail: flxj_lxl@163.com

Abstract. The air pollution problem has become a serious environmental problem facing many cities in China in recent years. In this paper, we introduced the Gradient Boosting Decision Tree (GBDT) method into the base learner training process of Bagging framework, and proposed a prediction model Bagging-GBDT based on Bagging ensemble learning framework. Based on this prediction model, we selected Beijing city of China as an example and established a PM2.5 concentration prediction model to forecast the PM2.5 concentration for the next 48 hours at a given time point. To measure the validity of the model, we also trained support vector machine regression models and random forest model to calculate three statistical indicators (RMSE, MAE and R²) for the proposed models on the test set to compare models performance. The experimental results show that our Bagging-GBDT model can better reduce the prediction bias and variance, and the prediction effect is better than SVR and random forest models.

1. Introduction
In recent years, with China's steady and rapid economic development, urban and rural industrialization construction have improved greatly. However, environmental protection lags far behind industrialization process. More and more cities have appeared serious air pollution problems. The study of this paper focused on the data-driven method for air pollutions prediction, with Beijing as an example. The main contributions include:(1)We built a mixed machine learning model called Bagging-GBDT based on ensemble learning for hourly PM2.5 concentration prediction. Base on the bias-variance decomposition of machine learning problems, this study draw Gradient Boosting Decision Tree (GBDT) into the training mode of the base learner of the Bagging framework to improve the prediction effect of the model. (2)We compared the mixed model with Random Forest (RF) model and support vector regression (SVR) model. The measurement indicators include the Root Mean Square Error (RMSE), the Mean Absolute Error (MAE) and the Coefficient of Determination (R²), which were computed on test set.

2. Related Work
Urban air pollution is a common environmental problem in many parts of the world. There are currently many research methods for the prediction of atmospheric pollutant concentration. One of the traditional approaches is physics-chemical methods based on numerical calculations. The framework mainly establishes a transformation equation between various physical and chemical components in the atmosphere and simulates air pollutants formation, accumulation, diffusion and migration.
processes by calculating the state transition relationship between various components. The method has high calculation cost, and the required meteorological data and pollutant data need to have high precision and finer time granularity and correctness. At the same time, due to the dramatic changes in the meteorological environment near the underlying surface, the establishment of such models involving a variety of physical and chemical processes is also over-reliant on relevant expertise.

Another alternative prediction method is based on the data-driven approach, which can be divided into machine learning regression models on account of statistical learning theory and various deep neural network models based on deep learning. Among the machine learning models commonly used in air quality prediction are support vector machine regression, multiple linear regression, and decision tree regression. Sabah et al. (2005) used both multiple linear and principal component regression methods to predict the concentration of ozone in the atmosphere [1]. Liu et al. (2016) studied to come up with a new model of collaborative forecasting using Support Vector Regression (SVR) for Urban Air Quality Index (AQI) prediction in China [2]. Jana et al. (2017) develop a Bayesian approach to forecasting PM10 and O3 levels that efficiently deals with extensive amounts of input parameters [3]. But there are also many problems with these commonly used models. Multiple linear regression methods lack the ability to deal with nonlinear relationships between data, which making it difficult to express the potential correlations of various physical and chemical components in the atmosphere. Support vector machines are subject to the choice of kernel functions. Decision tree regression requires control parameters to achieve a balance between under-fitting and over-fitting. At the same time, these models need to solve the problem of feature selection before training. The prediction effect has a lot to do with the choice of features. It is well known that feature engineering is complex and depends on professional knowledge.

Ensemble learning models, Compared to single machine learning models such as decision trees or SVR, have shortcomings in prediction accuracy. It combines a family of base learners into a unified framework for prediction through different combination strategies [4,5,6]. Usually and the effect is better than a single model. For air quality prediction, Stuart K. Grange et al. (2017) develop a random forest model to estimate daily PM10 concentrations [7]. Zhang et al. (2017) proposed a multi-channel ensemble learning framework, which based on Deep Boltzmann Machine sufficient statistics feature mapping [8].

With the advancement of computer computing power and the development of related theories in recent years, deep learning on the deep neural network has made great progress. The algorithm based on the deep network model infiltrates into various data-intensive research fields and achieves good results. S.A. Abdul-Wahab et al. (2002) proposed a neural network for ozone modelling in the lower atmosphere [9]. Li et al. (2016) proposed a novel spatiotemporal deep learning (STDL)-based air quality prediction method that inherently considers spatial and temporal correlations [10]. Fan et al. (2017) proposed a spatiotemporal prediction framework based on missing value processing algorithms and deep recurrent neural network (DRNN) to predict air pollution using missing time series data with missing value [11]. Li et al. (2017) proposed a novel long short-term memory neural network extended (LSTME) model that inherently considers spatiotemporal correlations is proposed for air pollutant concentration prediction [12]. Athira V et al. (2018) studied Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) to forecast air pollutions based on the pollution and meteorological data [13]. The neural network model is superior to the traditional machine learning model in prediction accuracy, and deep learning introduces the concept of feature learning, which generates good features through the network itself. Deep neural networks tend to fall into local minimum points during training. At the same time, the complex network structure requires more training data to solve the over-fitting problem. In addition, it is a black box model, and the algorithm established by it is poorly interpretable.

3. Methodology

In this section, we first briefly introduced the preparation of model construction from data processing and feature selection. The third part focused on the trade-offs of the model from the perspective of bias
3.1 Data
The research area of this paper is Beijing, China. As municipal construction, there are 35 air quality monitoring stations in Beijing. Our study collected hourly historical monitoring data from 35 air quality monitoring stations in Beijing from Beijing Municipal Environmental Monitoring Center website. The time interval is from January 1, 2017 to May 1, 2018. Site monitoring indicators include: PM2.5, PM10, CO, NO2, SO2 and O3, where CO unit is (mg/m³), other indicators are (micrograms/cubic meter), as shown in the Table 1.

At the same time, we collected observation data of 18 meteorological observation stations in Beijing from two website: caiyunapp.com and weather.com.cn (China weather website). The monitoring indicators include: weather, temperature, pressure, humidity, wind speed, wind direction.

3.2 Feature Engineering
In order to establish a PM2.5 regression prediction model, the paper selection features include:
1. Time characteristics: The current time belongs to the first few hours of the day and the days of the week, as well as the first few weeks of the year.
2. Meteorological characteristics: temperature, pressure, humidity, wind speed, and wind direction.
3. Monitor indicator characteristics: The concentration of major air pollutants, such as PM10 concentration.

3.3 Model Bias and Variance
Usually we want the learning algorithm to achieve lower errors in both the training set and the test set, especially focusing on the reduction of generalization errors. Bias-variance decomposition provides a tool to explain the generalization performance of learning algorithms. We know that the results of the algorithm learning on different training sets may be different. For sample x, let yD be the mark of x on the data set D, y be the true mark of x, and f(x, D) be the predicted value of x. Noise is defined as:

$$\epsilon^2 = E_D \left[ (y_D - y)^2 \right]$$

Then we get the generalization error formula:

$$E(f; D) = bias^2(x) + var(x) + \epsilon^2$$

That is, the generalization error based on the mean square error regression task can be decomposed into the sum of deviation, variance and noise. According to the definition, we know that the bias characterizes the degree of deviation between the expected prediction of the learning algorithm and the true value. The lower deviation indicates the ability of the algorithm to fit the training set data more strongly. The variance characterizes the degree of change in the algorithm performance, caused by changes in the same scale data set. That is to say, the variance expresses impact of data perturbations on the algorithm.

Considering the ensemble method from the perspective of bias-variance decomposition, Bagging mainly focuses on reducing the variance. Each base learner randomlysamples the returned samples from the original data set and then trains the learners on it. Finally these learners are combined into a single unit. The data perturbation is simulated by bootstrap sampling and other strategies, which
makes the base learners of Bagging to be diversity as large as possible. Its representative algorithm is random forest.

Boosting method focuses on reducing the bias. It first trains a base learner from the initial training set, then adjusts the training set sample distribution according to the performance of the base learner, iteratively trains a series of learners. And the samples, the model made mistakes in the last round of training, will receive more attention in the next round of training. Since the base learner obtained from each round of training corrects some errors of the previous model, Boosting is a training process that gradually reduces the deviation in a certain sense.

3.4 Method Construction
In order to reduce the bias of the model, our study considered Bagging as the main frame. It is noted that usually we select a single learning model as the base learner training algorithm in Bagging algorithm. In this paper, we introduced the Boosting strategy GDBT into the training of the base learner to improve the strength of the individual learner under the Bagging framework. Based on the preference of bias and variance for Bagging and Boosting, we constructed the algorithm framework is as follows:

Algorithm: Bagging-GBDT

| Input: | Original data set $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_S, y_S)\}$, Contains S samples; |
| Resampling ratio $\theta$; | |
| Number of training rounds $T$; | |
| Process: | For $t=1,2,3,\ldots,T$ do |
| | Bootstrap sampling to obtain a training set of size $N=\theta$ $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ |
| | Initialize $f^t_0(x) = \min_{\gamma} \sum_{i=1}^{N} L(y_i, \gamma)$ |
| For $m=1,2,3,\ldots,M$ do |
| For $i=1,2,3,\ldots,N$ compute |
| $r^t_{mi} = -\left[ \frac{\partial L(y_i, f^t(x_i))}{\partial f^t(x_i)} \right]_{f^t = f^t_{m-1}}$ |
| End for |
| Fit a CART to the target $r^t_{mi}$ giving terminal regions $R_{mj}, j=1,2,\ldots,J_m$ |
| For $j=1,2,\ldots,J_m$ compute |
| $\gamma_{mj} = \min_{\gamma} \sum_{x \in R_{mj}} L(y_i, f_{m-1}(x) + \gamma), f_{m-1} = f^t_{m-1}$ |
| Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{mj} I(x \in R_{mj}), f_m = f^t_m$ |
| End for |
| Output | $f^t(x) = f_M(x)$ |
| End for |
| Update $h_t(x) = h_{t-1}(x) + f^t(x)$ |
| End for |
| Output: | $H(x) = \sum_{t=1}^{T} h_t(x)$ |
4. Experiment and Analysis

In this section, we trained three regression prediction models: the support vector regression model, the random forest model, and the Bagging-GBDT hybrid model proposed in the previous section. The data used include hourly air monitoring historical data from 35 stations in Beijing from April 1, 2017 to April 1, 2018, as well as meteorological historical data at the same time. All stations were trained at the same time. After the training was completed, the RMSE, MAE, and R² values were calculated on the test set for each of the three models, and the performance of each model was compared.

For our hybrid method, we set the model key parameters to train two models: Bagging-GBDT-1 model corresponds to the number of training rounds of 5, the number of GBDT basic decision trees per round was 20, and the maximum height was set to 6. Bagging-GBDT-2 model corresponds to the number of training rounds was 10. The number of GBDT basic decision trees per round was 50, and the maximum height was set to 6. We used these five models to fit the training set and use the obtained model to measure the PM2.5 concentration of the air quality monitoring station for the next 48 hours.

As shown in Figure 1-4, we selected four urban monitoring stations, which respectively showed the results of PM2.5 concentration prediction for the site for 48 hours from May 30, 2018, 0:00 according to the five models.
We used the monitoring data from March 31, 2018 to June 1, 2018 as a test set to test the five models. The results of the calculation of the three evaluation indicators are as Table 2. It shows that our model can achieve better results than other models on the RMSE, MAE and R2 indicators. From the table we can see that the performance of the Bagging-GBDT-1 model is comparable to that of the radial basis kernel support vector regression model and is weaker than the random forest model. When we increase the number of training rounds and the number of GBDT basic decision trees, we can significantly improve the accuracy of the model and its ability to fit the data. Explain that our strategy can reduce the deviation and variance of the model.

| Model            | RMSE     | MAE      | R²        |
|------------------|----------|----------|-----------|
| LinearSVR        | 18.35375985 | 10.03436574 | 0.71259898 |
| RBFSVR           | 17.36231768 | 9.41763899  | 0.79980291 |
| RF               | 16.99717096 | 8.9511194  | 0.89525617 |
| Bagging-GBDT-1   | 17.54299084 | 9.66779886  | 0.79912286 |
| Bagging-GBDT-2   | 15.27241778 | 7.6977284  | 0.9563944  |

5. Conclusion
As one of the main pollutants in the air, PM2.5 is an important cause of smog weather. In view of its important harmful effects on the health of urban residents, it is of great theoretical and practical significance to study the PM2.5 concentration prediction problem, which can help the government and the public to formulate countermeasures.

Starting from the bias variance tradeoff of the prediction model, this paper introduced GBDT into the training of Bagging framework to improve base learner strength and then established the Bagging-GBDT model. And we used the air quality and meteorological historical data to train the PM2.5 prediction model. Comparative experiments show that the proposed method has higher relative prediction accuracy, the basic decision tree depth does not need to exceed the random forest component because of the Boosting strategy. Our method is superior in the test set and the generalization performance is better.

Through the research, it is found that the monitoring sites at different locations have obvious regional differences in the performance of PM2.5 concentration, that is, the concentration of PM2.5 is related to the spatial location, such as the existence of a certain potential pollution source near several sites. Therefore, further work will include: deeply researching the introduction of various spatial information to improve the prediction effect of the model; on the other hand, researching the temporal and spatial correlation of the prediction results of each site to explore the potential pollution patterns.

Acknowledgments
The paper is supported in part by the National Natural Science Foundation of China under Grant No. 61672022 and No. 61272036, Key Disciplines of Software Engineering of Shanghai Polytechnic University under Grant No. XKKZD1604, and Graduate Innovation Program No. A01GY17F022 and No. A01GY18F022-g13, and Shanghai Pujiang Program under Grant No.18PJ1433400.

References
[1] Sabah A. Abdul-Wahaba, Charles S. Bakheetb, Saleh M. Al-Alawi. Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. Environmental Modelling & Software 20 (2005), pp.1263–1271.
[2] Bing-Chun Liu, Ardhant Binaykia, Pei-Chann Chang, Manoj Kumar Tiwari, Cheng-Chin Tsao. Urban air quality forecasting based on multidimensional collaborative Support Vector Regression (SVR): A case study of Beijing-Tianjin-Shijiazhuang. PLoS ONE, Volume 12, Issue 7, July 2017, e0179763.
[3] Faganeli Pucer, J., Pirš, G. & Štrumbelj, E. A Bayesian approach to forecasting daily air-pollutant levels. Knowledge Information System (2018), pp.1--20.

[4] Thomas G. Dietterich. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. Machine Learning, 40, pp.139--157.

[5] Ali, K. M. & Pazzani, M. J. Error reduction through learning multiple descriptions. Machine Learning, 1996., 24(3), pp.173--202.

[6] Valentini, Giorgio and Thomas G. Dietterich. “Low Bias Bagged Support Vector Machines.” ICML (2003), pp.752-759.

[7] Stuart K. Grange, David C. Carslaw, Alastair C. Lewis, Eirini Boleti, and Christoph Hueglin. Random forest meteorological normalisation models for Swiss PM10 trend analysis. Atmospheric Chemistry and Physics , 18(9), pp. 6223-6239.

[8] Chao Zhang, Junchi Yan, Yunting Li, Feng Sun, Jinghai Yan, Dawei Zhang, Xiaoguang Rui, Rongfang Bie. Early Air Pollution Forecasting as a Service: an Ensemble Learning Approach. 2017 IEEE 24th International Conference on Web Services, pp. 636-643.

[9] S.A. Abdul-Wahab, S.M. Al-Alawi. Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks. Environmental Modelling & Software. 2002, 17(3), pp.219--228.

[10] Xiang Li, Ling Peng, Yuan Hu, Jing Shao, Tianhe Chi. Deep learning architecture for air quality predictions. Environmental Science and Pollution Research. Volume 23, Issue 22, 1 November 2016, Pages 22408--22417.

[11] Junxiang Fan, Qi Li, Junxiong Hou, Xiao Feng, Hamed Karimian, Shaofu Lin. A Spatiotemporal Prediction Framework for Air Pollution Based on Deep RNN.ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Volume 4, Issue 4W2, 19 October 2013, Pages 15--22.

[12] Xiang Li, Ling Peng, Xiaojing Yao, Shaolong Cui, Yuan Hu, Chengzeng You, Tianhe Chi. Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. Environmental Pollution. Volume 231, 2017, Pages 997--1004.

[13] Athira V, Geetha P, Vinayakumar Rab, Soman K P. DeepAirNet: Applying Recurrent Networks for Air Quality Prediction. Procedia Computer Science. Volume 132, 2018, Pages 1394--1403.