EWA-GBDT: A Novel Air Quality Prediction Model Combining Exponentially Weighted Averages and Gradient Boosting Decision Tree

Jibing Gong, Dan Wang, Da Chen, and Shuli Wang

***Abstract*—Air quality prediction is a hot topic in the field of meteorology. Challenges still exist following consideration of the uncertainty of atmospheric pollutant emission sources, as well as the multi-dimensional, multi-scale, and non-stationary characteristics of meteorological environment data. For example, traditional statistical forecasting methods usually fit a nonlinear relationship between meteorological features and pollutants, which is why it is extremely difficult to learn their models.**

To address these challenges, we propose a novel air quality prediction model combining Exponentially Weighted Averages and Gradient Boosting Decision Tree (EWA-GBDT). More specifically, we first collected two real-world datasets, including 1) the daily concentration data of six pollutants from 01/01/2014 to 31/12/2016 and 2) the daily concentration data of meteorological features in the cities over Shijiazhuang and Xingtai from 01/01/2014 to 02/28/2017. Then, we extracted 13 types of meteorological features using the Support Vector Machine Recursive Feature Elimination method. From the respective of pollutant concentration, these features are the highest correlated with each other. Next, we applied the EWA principle to compute the above features and pollutant concentration for obtaining meteorological expectation values. Finally, considering the excellent overall prediction performance of ensemble learning, we utilized its GBDT algorithm to predict the concentration value of pollutants and thus output the air quality level. We conducted experiments on the two datasets, and the results demonstrate that EWA-GBDT outperforms other baseline methods in terms of root-mean-square deviation, mean absolute error, and the square of R.

***Index Terms*—Air quality prediction, exponentially weighted averages, gradient boosting decision tree, recursive feature elimination.**

I. Introduction

Real-time air quality information, accompanying the rapid industrialization and urbanization, air pollution has become increasingly severe in our country, and air quality has formed as a key environmental issue of national and social concerns at a fine spatiotemporal scale [1]. Studies have shown that meteorological features have become a dominant factor affecting urban air pollution [2]. Among various air pollutants, particulate matters are one of the deadliest and difficult to get rid of by the immune system once inhaled. Therefore, it is of great importance to accurately monitor and predict $PM_{2.5}$ and other pollutants.

The methods of air quality prediction are becoming more and more widespread. Research in machine learning has developed rapidly in recent years. Existing methods mainly focus on machine learning models such as time series analysis, multiple regression model, artificial neural network model, etc., but these models make more assumptions about air pollution changes, which could lead to a large amount of training data unavailable, moreover, its main prediction period within 8 to 24 hours. Plus, the generalization ability of neural networks is minimal. Therefore, by studying the correlation between meteorological features and atmospheric pollutants, predicting air quality is a key issue for protecting human health and effectively controlling air pollution.

The mainstream methodology of air quality prediction is to combine the air quality model with statistical methods, which has the characteristics of weak regularity, instability, and easy mutation. Its time series and geographic topological relationship of modeling pose challenges to machine learning: 1) how to identify discriminative meteorological features from a variety of data sources and incorporate it consider their impact on air quality, 2) how to excavate the non-linear relationship between meteorological features and pollutants, and 3) how to integrate major pollutants and meteorological features in different cities and seasons.

Aiming at the main defects of the current air quality prediction model, the recursive feature elimination (RFE) is first leveraged to retrieve the non-linear relationship between meteorological features and different pollutants to select the first 13 meteorological features with strong concentration of each pollutant, moreover, utilize the exponentially weighted and respectively averaged with a weight value of 0.9 to extract data trends. Specifically, the meteorological value(the values of these 13 meteorological features and pollutants) of the first 10 days plus 0.1 times of the weather value of the same day as the expected meteorological value, that is, the values of meteorological features and pollutants, are exponentially weighted averaged, which makes the data smoother and less noisy. Finally, processed results as the input of the Gradient Boosting Decision Tree (GBDT) in ensemble learning (EL), which leverages multiple basic learners to train and integrate the learning results of multiple basic learners. Hence, it is effect better than a single learner more convincing.

To address this issue, the main contributions of this study lie in the following three aspects:

1) The Support Vector Machine (SVM)-RFE method was

Manuscript received April 29, 2019; revised July 1, 2019. This work was supported by the Hebei Natural Science Foundation of China (Grant No. F2015203280) and the Graduate Education Teaching Reform Project of Yanshan University (Grant No. JG201616).

The authors are with the School of Information Science and Engineering, Yanshan University, Qinhuangdao 066044, China (e-mail: gongjibing@163.com , wangdanysu8100@163.com, chenda_ysu@163.com wangshuli@ysu.edu.cn).

DOI: 10.7763/IJMO.2019.V9.706

used to select the meteorological features with a good correlation with pollutants. Owing to the fact that pollutants affecting cities in different quarters are different, the meteorological features affecting the concentration of pollutants in different cities in different seasons are also different.

2) Using Exponentially Weighted Averages (EWA)-GBDT, we also quantified the impact on air quality from a variety of temporal factors, exponentially weighting the meteorological features concentration, establishing the pollutant concentration prediction model in different seasons.
3) We evaluated our approach with different competing baselines in different metrics, proving the accuracy and efficiency of the EWA-GBDT. The results show the advantages of our approach over for baselines.

The remainder of our paper is organized as follows: Section 1 presents the introduction. Section II describes the theoretical basis and process of this algorithm. Section III is the experimental settings and results analysis in detail. Section IV introduces the related research progress of air quality prediction models. Section 5 concludes the paper with a discussion of future work.

The symbols and meanings utilized in this paper are shown in Table I.

TABLE I: THE ARRANGEMENT OF CHANNELS

Symbol	Description
R	Feature sorted list
n	Sample size
ri	Correlation coefficient between characteristic i and the dependent variable
L	Loss function
n_estimators	The number of gradient boosting decision tree
max_depth	Maximum depth of individual regression estimator
learning_rate	Learning rate

II. CONSTRUCTION OF AIR QUALITY PREDICT MODEL

A. Framework

In this study, we present a EWA-GBDT method to predict the concentration of pollutants. Firstly, a feature selection based on the SVM-RFE was utilized to select meteorological features of good correlation with pollutants. After that, EWA-GBDT was assigned to predict the concentration of pollutants to predict air quality.

We adopted the model according to the framework of co-training and the philosophy shown in Fig.1. It mainly consists of two parts: 1) feature selection, using SVM-RFE, calculating the correlation score between all meteorological features and each kind of pollutant, respectively eliminating the meteorological factor with the lowest score, and repeating until the number of remaining meteorological features reaches the required number and 2) the model predicts that the selected meteorological features and the values of the pollutants are weighted and averaged, and the results are converted into a consistent input to model the GBDT to predict the concentration value of the pollutants. As depicted in Fig. 1, our framework consists of two major parts: the feature selection and the pollutant concentration prediction.

Feature selection process: In this stage, we utilized SVM-RFE for feature selection, which is a sequence reverse selection algorithm based on the SVM maximum interval principle. First, the original feature set is initialized; second, the relevant scores of each meteorological feature with pollutants are generated, and the minimum score of the meteorological features is eliminated. The retrieving samples and feature sets are updated, and the remaining meteorological features are iterated until its size is up to zero. In Fig. 1, the upper frame represents this process where *f* is the meteorological features of the input.

Pollutant concentration prediction process: In this flow (represented by the lower frame), we leverage the EWA-GBDT to construct the pollutant concentration prediction model. First, the Exponentially Weighted Averages is assigned to optimize the data of meteorological features and pollutant concentration. Next, the GBDT is used to replace the residual of pollutant concentration with the loss function in the gradient direction, and the regression prediction retrieving of meteorological features and daily pollutant concentration value is carried out.

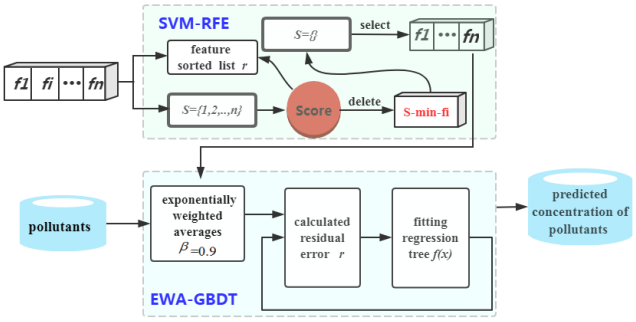


Fig. 1. The architecture of EWA-GBDT.

B. SVM-RFE

In this section, our goal is to identify high correlation meteorological features with each pollutant concentration. RFE works iteratively to train the basic model [3] (such as SVM, decision tree, *et al.*); features are scored according to the weights of each feature, which are captured from SVM. The best or worst features are selected for elimination after each training, and the remaining features are trained until all the features are traversed, or the number of remaining features reaches the required number [4].

In practice, SVM is assigned as the infrastructure learner, SVM-RFE integrating SVM and RFE. Perform feature selection by recursively reducing the size of the feature set based on SVM regression. First, the current feature index sequence S is initialized with all features, new sample captured according to the current features, and second, SVM training is performed to get weight ω, and then calculate the sorting score c_i , find the feature with the smallest score and removed from S and also added to the feature sorted list r in order, finally， the process is repeated continuously until S is empty or the number of features in r reaches the desired level. The meteorological features, which have little correlation with pollutant concentration, are eliminated to achieve the goal of reducing dimension and calculation and improving prediction accuracy through feature selection. Algorithm 1 outlines the SVM-RFE training process. Here, S is the current feature index sequence, r represents a feature-sorted list, ω is the weight vector of SVM, c_i is the ranking score of features,

and *p* represents the current sort score minimum.

The SVM-RFE algorithm traverses all the feature index sequence *S* in each experiment; the traversal times are *n*, so the average time complexity is $O(n)$, and the space complexity is $O(1)$.

Algorithm1: SVM-RFE

Input: train dataset *T*; $T = \{x_i, y_i\}_{i=1}^{N}$
Output: feature-sorted list *r*
Steps:
1. initialize original feature set $S = \{1,2, \dots, n\}$,*r* = []
2. for *i* in *S* do
3. obtain new training set *X* from the candidate feature set
4. use $\min \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i \alpha_j y_i y_j (x_i * x_j) - \sum_{i=1}^{N}\alpha_i$ training SVM learner to get ω
5. calculate the ranking criteria score using the equation:
$$c_i = \omega_i^2 \ , i = 1,2, \dots, n$$
6. find the features with the smallest ranking score
$$p = argminc_k$$
7. update feature sets *r = {S(p), r}*
8. eliminate this feature in *S:S = S/S(p)*
9. if *S*= [ ]
10. end for

C. EWA-GBDT

The nature of the EWA is the exponentially weighted moving average [5]. The weight of each value decreases exponentially with time; the closer the time is, the greater the weight is, while the older data is also given a certain weight. At time *t*, V_t can be captured from the actual observations: $V_t = \beta * V_{t-1} + (1-\beta) * \theta_t$, here, V_t means the estimated value at time *t* (e.g., the EWA). θ_t is the measured value at time *t*. *n* is the total time observed. $\beta (0 < \beta < 1$ represents the weighting coefficient for historical measurements. In each element V_n of the new sequence, $V_1, V_2, V_3, \dots, V_t, V_n$ can be approximated as the weighted average of the first 11-β elements until the original sequence ends to the *t-th* element. The closer the coefficient β is 1, the higher the weight of the current sample value. The lower the weight of the past measured value, the stronger the timeliness of the estimated value, and vice versa. When β is 0.9, V represents the weighted average of over 10 days. If we set the β value to be 0.98, then we are calculating the EWA of over 50 days. In practice, we assigned the exponentially weighted and averaged with a weight value of 0.9 to extract data trends. The weather value of 10 days before a certain day adds 0.1 times of the weather value of the day as the expected weather value, making the data smoother and less noisy.

The GBDT is one of boosting algorithm in EL [6], and it is also an improvement of boosting algorithm. After the end of each step, the original boosting algorithm increases the weight of the wrong classification points and reduces the weight of the correct classification points [7]. This makes it possible for some points to be "seriously concerned" if they are previously misclassified for a long time; that is to say, they are given a high weight. The smaller the base classifier weight value is, the smaller the error rate is. The smaller the error rate is, the larger the weight value of the base classifier is. After *n* iterations, *n* simple basic learners will be obtained. We continue to merge the outputs together and choose the one that has the greatest votes. The core of GBDT is that each calculation is done to reduce the residuals of the previous one [8]. Every new model is established in the gradient direction of the residual reduction. In GBDT, each new model is established to reduce the residue of the previous model to the gradient direction, and the negative gradient of the loss function in the current model is used as the approximation of the residual of the tree algorithm in the regression process. After that, a regression tree is fitted with the residue as the output. This is a great difference from the traditional boosting algorithm for weighting correct and incorrect samples.

The EWA-GBDT algorithm process is reproduced below.

1) Selection of characteristic meteorological features requiring exponential weighted averaging according to actual conditions.
2) The EWA processing of the corresponding meteorological features data from the selection result of the previous step. $V_t = \beta * V_{t-1} + (1-\beta) * \theta_t$, where V_t represents the average of day *t* and θ_t represents the raw data value of day *t*. β represents an adjustable hyper parameter value.
3) Initialize, estimate the constant value that minimizes the loss function, refer to (1), where $f_0(x)$ means the initial GBDT, L is the loss function.
$$f_o(x) = argmin_\rho \sum_{i=1}^{N} L(y_i, \rho) \quad (1)$$
4) Iteratively generate M regression trees.
   - for *i* = 1 to *n*, calculate the negative gradient value of the loss function and use it as an estimate of the residual r_{im}
   - fit a regression tree $f(x)$ to the residual r_{im}
   - calculate the ρ_m of the gradient descent
   - update $f_m(x)$
   - output model $f_M(x)$

Algorithm2: EWA-GBDT

Input: train dataset *T*; $T = (x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, loss function *L(y, f(x))*
Output: gradient boosting decision tree *f(x)*
Step:
1. select the characteristic meteorological features that need to apply EWA
2. EWA processing of corresponding meteorological factor data
$$V_t = \beta * V_{t-1} + (1-\beta) * \theta_t$$
3. initialization: for *m = 1* to *M* do
$$r_{im} = -\left[\frac{\partial L(y_i, f_{m-1}(x_i))}{\partial f_{m-1}(x_i)}\right] f(x) = f_{m-1}(x_i)$$
4. for *i = 1* to *N* calculating residuals
5. the *m-th* regression tree *g* was trained, and the area divided by its leaf node was $R_{m,j}$
6. for each leaf node of regression tree g, calculate the step length of gradient descent
$$\rho_m = \text{argmin}_\rho \sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + \rho g_m(x_i))$$
7.update
$$f_m(x) = f_{m-1}(x) + lr * \rho_m g_m(x)$$
8.end for
9.end for
10. get the final gradient boosting decision tree
$$f_M(x) = \sum_{m=1}^{M}\sum_{j=1}^{J} \rho_m g(x)$$

The concrete algorithm is presented in Algorithm 2. The EWA-GBDT algorithm fits a regression tree according to the negative gradient of loss function when traversing all samples of the training set. Because each leaf node of each tree generates ρ_m, the average time complexity of the

algorithm is $O(n^2)$, and the space complexity is $O(1)$.

III. Experimental Results and Analysis

A. Datasets

We utilized two cities, Shijiazhuang and Xingtai, as representative cities in the central and southern regions of Hebei Province. In addition, in practice, we first collect two real-word datasets, including 1) the daily concentration data of six pollutants from 01/01/2014 to 31/12/2016 and 2) the daily concentration data of meteorological features over Shijiazhuang and Xingtai from 01/01/2014 to 02/28/2017. The data is afterward preprocessed in the experiments, including noise reduction and null processing. For the two characteristics of surface ventilation coefficient and mixed layer height, some data have missing values. Accordingly, in the time direction, three nearest neighbors are selected to construct Lagrangian quadratic polynomials and further solve them. The value of the Lagrangian quadratic polynomials is the missing value of filling. After that, the data is divided into four parts according to city and quarter, which are the meteorological features and concentration data of different cities in the first quarter, second quarter, third quarter, and fourth quarter.

B. Evaluation Metrics

In order to evaluate the performance of our approach, we adopted the following metrics: *MAE*, *RMSE*, and R^2.

MAE. Mean absolute error; measures the absolute difference between two continuous variables, it is defined as

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{2}$$

RMSE. Indicates the Root-mean-square deviation, measures the differences between the predicted values and the actual values, it is defined as

$$RMSE = \frac{1}{n}\sqrt{\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|} \tag{3}$$

R^2**.** R-squared, which represents the ratio of the residual sum of squares to the total sum of squares. It is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}_i\right)^2}{\sum_{i=1}^{n}\left(\bar{y}_i - \hat{y}_i\right)^2} \tag{4}$$

where n denotes the size of training set, y_i represents the true value, $\hat{y}_i$ represents the predicted value, and $\bar{y}_i$ means the mean of the true value.

C. Baseline Methods

To evaluate our approach, we compare our approach with the following four baseline methods:

SVM. SVM algorithm is used for regression prediction, and pollutant concentration is predicted based on historical data.

RF. Random forest algorithm is used for regression prediction and pollutant concentration prediction based on historical data.

GBDT. A simple algorithm used to predict the model by regression, and the pollutant concentration is predicted by historical data.

EWA-GBDT-1. Only some closely related meteorological features (highest temperature, static stability index, average temperature, and minimum relative humidity) are applied to the EWA model.

D. Parameter Setting

We conducted the EWA-GBDT using the scikit-learn in Python. For our model, we first used the RFE based on the SVM regression method, with kernel as linear as the basic learner. Operates the number of features on 13 according to actual condition, after that, 13 meteorological features that have a strong correlation with each pollutant concentration are selected. Specifically, our approach utilizes the extract result in the next step to exponentially weighted average, where we set the β to 0.9. Finally, the processed value is regarded as the input of the EWA-GBDT, and the training set and the test set are split in a ratio of 4:1. The optimal parameters are selected at this time: *n_estimators* is set to 400, *max_depth* is set to 4, *min_samples_split* is 2, and *learning_rate* is 0.01. For the other baseline methods, we use 80% of the training data to optimize their parameters. All experiments were performed on one machine with an Intel (R) Core i5-8250U (Main frequency 3.40 GHz) and a RAM of 8 GB.

E. Analysis of Prediction Performance

a) SVM-RFE performance

The correlation analysis of meteorological features with pollutants based on SVM-RFE solves the association degree of meteorological features and pollutants in different cities and seasons. Convert each meteorological feature into a consistent input, likewise utilizing a single pollutant as the target analysis object to establish an RFE analysis model for each meteorological features and pollutant concentration, and ranking the correlation between each meteorological feature with pollutant concentration.

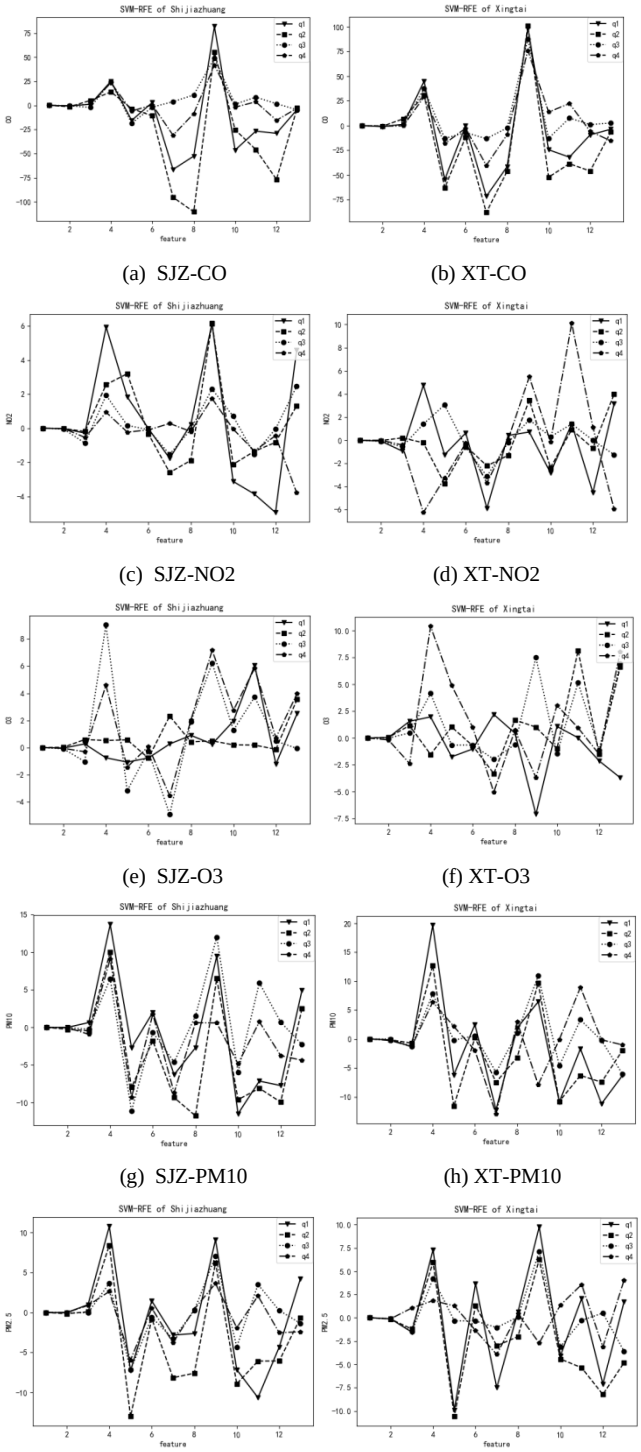
TABLE II: The X-Coordinate Corresponds to the Actual Meteorological Features

Serial number	Meteorological features	Serial number	Meteorological features
1	Surface ventilation coefficient	8	Station pressure
2	Mixed layer height	9	Static stability index
3	Relative humidity	10	Sunshine hours
4	Maximum temperature	11	Average temperature
5	Daily maximum wind speed	12	Precipitation
6	Minimum relative humidity	13	Average wind speed
7	Minimum temperature		

For visualization, the SVM-REF results for diverse cities is demonstrated in Fig. 2. It shows the relationship between meteorological features and pollutant concentration, where a) represents Shijiazhuang and b) represents Xingtai. The x-coordinate represents each meteorological features, and the y-coordinate indicates the degree of correlation with each pollutant. The top 13 features ranked by importance are shown in Table II. In short, these features are very

discriminative.

From Fig. 2, cities with different meteorological features show different impacts varying by quarters. A surprising discovery is that meteorological features, which has the greatest impact on *CO* concentration, is static stability index; the meteorological feature that has the greatest impact on NO_2 concentration is static stability index. Xingtai is the average temperature; the most significant influence on the concentration of O_3 in Shijiazhuang is the static stability index, Xingtai is the maximum temperature; the maximum temperature is most affected by the PM_{10} concentration; the highest affecting the $PM_{2.5}$ concentration in Shijiazhuang is the maximum temperature, Xingtai is the static stability index; static stability index is the most important factor affecting SO_2 concentration both in Shijiazhuang and Xingtai.



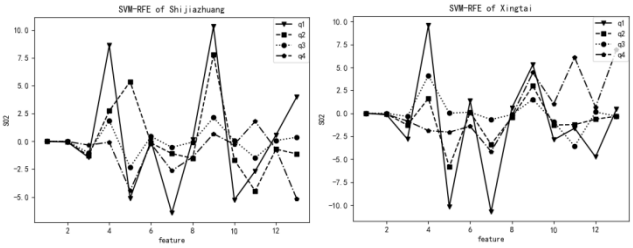
(a) SJZ-CO (b) XT-CO

(c) SJZ-NO2 (d) XT-NO2

(e) SJZ-O3 (f) XT-O3

(g) SJZ-PM10 (h) XT-PM10

(i) SJZ-PM2.5 (j) XT-PM2.5



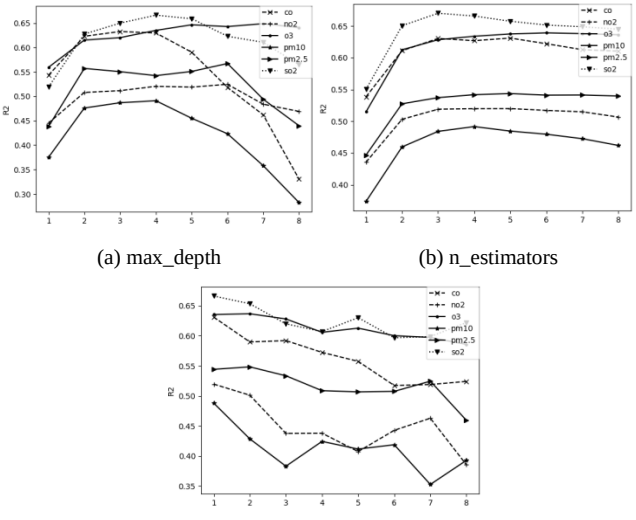
(k) SJZ-SO2 (l) XT-SO2

Fig. 2. The relationship between pollutant concentration and meteorological features.

b) EWA-GBDT performance

On the basis of SVM-RFE, 13 meteorological features that have a strong correlation with pollutant concentration were identified, and they were exponentially weighted averaged with pollutant concentrations in different cities, where β was 0.9, and each meteorological feature serves as an input feature corresponding to a set of vector values. The GBDT prediction model for each pollutant concentration is established according to different quarters.

In order to improve the performances of our model, the model parameters require to be adjusted. The parameters selected in this study are *max_depth*, *n_estimators*, and *learning_rate*. The selected evaluation metrics are R^2, *MAE*, and *RMSE*. Take Shijiazhuang four quarters as an example, adjust the parameters were evaluated using the regression model R^2 to obtain the experimental parameter results, which are shown in Fig. 3.



(a) max_depth (b) n_estimators

(c) learning_rate

Fig. 3. The influence of parameters.

As illustrated in Fig. 3, the y-coordinate indicates the regression evaluation metrics R^2, where a) the x-coordinate of the subfigure represents the value of the *max_depth* parameter. b) the x-coordinate of the subfigure represents the value of the *n_estimators* parameter, and c) the x-coordinate of the subfigure represents the *learning_rate* and one small square represents 100 times. The values of the parameters, the x-coordinate 1 to 8 correspond to the values: 0.01, 0.03, 0.06, 0.09, 0.12, 0.15, 0.18, and 0.21.

Obtained through analytical experiments: in subfigure (a),

the model prediction performance is the best when the parameter *max_depth* is 4, and the GBDT can control the number of nodes in the tree according to the specific problem to achieve the best prediction effect in subfigure(b). When the parameter *n_estimators* is 300 or 400, the predict effect is best in the prediction of each pollutant concentration. When *n_estimators* set 100 or 200, the training effect is inadequate, while the effect of continuing to increase is not significantly improved, indicating that the model is better. Over-fitting and under-fitting problems are solved, and in subfigure (c), when the parameter *learning_rate* is 0.01, the effect of improving the *learning_rate* is reduced. According to the results of the parameter adjustment, the optimal parameters are selected in this model, *n_estimators*: 400, *max_depth*: 4, *learning_rate*: 0.01.

The prediction results of $PM_{2.5}$ concentrations in Shijiazhuang and Xingtai are shown in Table III.

TABLE III: PREDICTION PERFORMANCE OF $PM_{2.5}$

	MAE	*RMSE*	R^2
Shijiazhuang	11.8878	15.8398	0.9172
Xingtai	10.4759	13.6248	0.9244

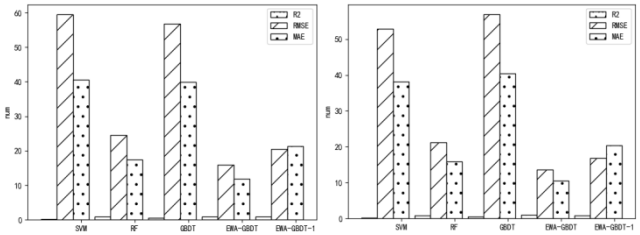
The concentration of $PM_{2.5}$ in the four quarters of Shijiazhuang is predicted. The results are shown in Table IV.

TABLE IV: FORECAST INDICATOR TABLE IN QUARTERLY $PM_{2.5}$ CONCENTRATION

	First quarter	Second quarter	Third quarter	Fourth quarter
MAE	8.7124	3.8924	6.4089	10.1090
RMSE	11.2899	4.7512	9.0184	13.2446
R^2	0.9490	0.927	0.8144	0.9543

The smaller the *RMSE*, the better the prediction effect of the model. From the perspective of *RMSE*, the model has the best prediction effect in the second and third quarters, while the first and fourth quarters are relatively inferior. It is actually very reasonable that maybe caused by uncontrollable external factors such as heating in winter, and high pollution events generally occur in winter.

As depicted in Fig. 4, for different cities, the EWA-GBDT model has stable performance compared with other baselines and brings a significant improvement on the prediction accuracy. The three evaluation metrics of our approach are achieving the best performance. This is very intuitive to understand that RF and GBDT have similar effects and are much more improved than the SVM. The EWA-GBDT-1 with partial EWA of some meteorological features has a slightly better performance compared with the RF and GBDT, further indicating that the model effect will be significant when the EWA is applied to all meteorological features.

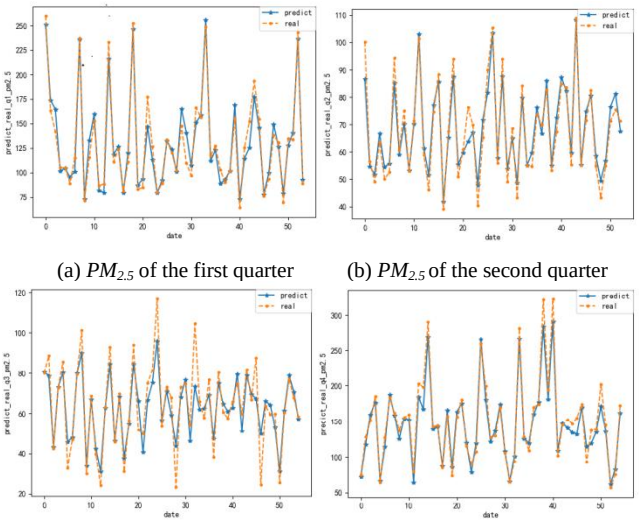


(a) Shijiazhuang (b) Xingtai

Fig. 4. Comparison with baseline methods.

Temporally speaking, all strongly correlated meteorological features (e.g., relative humidity, static stability index, and average temperature, *et al.*) extracted from SVM-RFE are weighted exponentially with the concentration of six pollutants. In other words, the daily concentration is expressed by the concentration average of 10 days before and after, which reduces the dimension of features and the sparsity of data.

We partitioned the data into four quarters. The concentration of $PM_{2.5}$ pollutants in the four quarters of Shijiazhuang is predicted. The results are shown in Fig. 5.



(a) $PM_{2.5}$ of the first quarter (b) $PM_{2.5}$ of the second quarter

(c) $PM_{2.5}$ of the third quarter (d) $PM_{2.5}$ of the fourth quarter

Fig. 5. Prediction results.

In Fig. 5, the x-coordinate indicates the date and the ordinate represents the concentration of $PM_{2.5}$. All these results well demonstrate the similarity predicted result of the model between the trends of the real. Clearly, every result fall in the around region of each real data, which verifies the validity of our method. In general, the GBDT is a regression algorithm in the direction of the gradient, which can quickly reduce the concentration difference between the predicted value and the real value, and reduce the time complexity of the algorithm, thus, making the model achieve better performance.

F. Discussion

This section further discusses the architecture we established, and provide an explanation of the previous analysis results.

1) Aligning with the co-training framework, our proposed method significantly outperforms the baseline methods. From Fig. 4, as expected, we can then infer that EWA-GBDT achieves the best performance. This indicates that our presented approach has a good performance in air quality prediction.
2) As shown in Fig. 2, compared with other baselines, RFE based on SVM enables the highlights of the correlation between meteorological features and pollutant concentration.
3) Clearly, Fig. 5 further reveals that EWA-GBDT tends to significantly improve the accuracy of pollutant concentration prediction. Compared with other baselines, GBDT is a regression algorithm in the direction of

gradient, which can quickly reduce the concentration difference between the predicted value and the real value, and reduce the time complexity of the algorithm, thus, making the model achieve better performance.

4) Further analysis shows that seasonal factors have a significant impact on the experimental results in practice. It is more effective in the second and third quarters than in the first and fourth quarters. Some reasonable explanations can help illustrate the consequences: heating in winter and high pollution events generally occur in winter and so on. Additionally, we will further study the first and fourth quarter, adding time series to achieve better results.

IV. Related Work

Air quality prediction methods mainly fall into two categories: classical dispersion models and data-driven models. Existing work typical for air quality prediction is conducted through neural network. Li X. C. [9] *et al*. extracted the uptrend intervals and calculated the causal strengths among spatially distributed sensors from the perspective of detecting pollution sources and mining pollution propagation patterns, used causal strengths to model the spatio-temporal uptrend events, afterward constructed causality graphs to determine pollution sources and propagation patterns. Zheng Y. [10] *et al.* proposed a semi-supervised learning approach based on a co-training framework that consists of two separated classifiers, the artificial neural network and the linear-chain conditional random field. Ibrahim K. [11] *et al.* proposed a novel model based on Long Short Term Memory networks to predict future values of air quality in a smart city. On the basis of traditional BP neural network, Li L. [12] *et al.* proposed a sample optimization method based on meteorological similarity criterion, which improves the prediction accuracy of the traditional neural network. Wu C. L. [13] *et al.* presented a representative method of regional space–time data and a $PM_{2.5}$ prediction model based on the convolution neural network. However, the slow convergence, easy to fall into local optimum, and poor robustness of the neural network also limit its further development.

Keller J. P. [14] *et al.* estimated time trends from an observed time series and used spatial smoothing methods to borrow strength between observations. Applied a spatio-temporal model that included a long-term spatial mean, time trends with spatially varying coefficients, and a spatio-temporal residual in each region. Feng X. [15] *et al.* proposed a novel hybrid model combining air mass trajectory analysis and wavelet transformation. Wang L. M. [16] *et al.* according to the distance correlation coefficient, the daily rolling statistical prediction of $PM_{2.5}$ concentration was carried out by using SVM regression algorithm. However, these methods all have one key issue: feature selection, which generally uses simple mutual information, Pearson correlation coefficient, distance correlation coefficient, etc. It is easy to fall into local optimum, with high uncertainty, no obvious effect, and not suitable for large data sets.

The GBDT plays a significant role in many fields. Kim T. K. [17] *et al.* proposed a method to speed up the evaluation time of a GBDT for fast-moving object tracking and segmentation problems. Hu J. F. [18] verified that the GBDT-based method may assist in the detection of driver fatigue. Pham H. D. [19] *et al.* applied a static malware detection method by Portable Executable analysis and GBDT algorithm. Additionally, Gong J. B. [20] *et al.* designed Friend++, a hybrid multi-individual recommendation model that integrates a weighted average method into the random walk framework by seamlessly employing social ties, behavior context, and personal information. References [21], [22] construct neural network models to mine deep information between features respectively from emotional analysis and semantics of social relationships.

V. Conclusion

For simultaneously capturing the factors affecting future air quality, we identified 13 meteorological features that have a strong correlation with pollutant concentration through RFE. Next, concentration of meteorological features are fed into EWA-GBDT to predict pollutant concentration and further study air quality. From this study, we obtained the following conclusions.

1) Based on the RFE of SVM, considering the non-linear relationship between meteorological features and pollutant concentration, moreover, extracting 13 meteorological features (e.g., stationary index, average temperature, surface ventilation coefficient, *et al.*) with a strong correlation with pollutant concentration.
2) This study has significantly improved the prediction that the accuracy of urban pollutant concentration relies on our selected input. Taking the concentration of $PM_{2.5}$ in Xingtai as an example, the RMSE value decreased to 13.6248 and the R^2 value increased to 0.9244.
3) This study has fully considered the influence of time factors on the prediction results by dividing the quarter, and improved the prediction accuracy. The model's forecasting effect in the second and third quarters is preferable than that in the first and fourth quarters, indicating that the model's improvement in summer is more striking than in winter. The model effect is accurate and stable compared to other traditional methods, which can provide a reference for operational forecasting. In addition, it has a finer spatial and temporal granularity.

Future work is to integrate the time series model. It is planned to use the historical data of the same month for the time series seasonal analysis to eliminate the instability of the data and improve the prediction accuracy.

References

[1] J. S. Wang and G. J. Song, "A deep spatial-temporal ensemble model for air quality prediction," *Neuro Computing*, vol. 314, no. 7, pp. 198-206, November 2018.

[2] Z. Y. Chen, B. Xu, and J. Cai, "Understanding temporal patterns and characteristics of air quality in Beijing: A local and regional perspective," *Atmospheric Environment*, vol. 127, pp. 303-315, February 2016.

[3] A. Filali, C. Jlass, and N. Arous, "Recursive feature elimination with ensemble learning using SOM," *International Journal of Computational Intelligence and Applications*, vol. 16, January 2017.

[4] G. S. Chen and Q. Z. Zheng, "Online chatter detection of the end milling based on wavelet packet transform and support vector machine recursive feature elimination," *International Journal of Advanced Manufacturing Technology*, vol. 95, pp. 775-784, March 1, 2018.

[5] H. Abdul, J. Brown, and M. Elena, "A new maximum exponentially weighted moving average control chart for monitoring process mean and dispersion," *Quality and Reliability Engineering International*, vol. 31, no. 8 , pp. 1587-1610, December 2015.
[6] H. D. Cui, D. L. Huang, and Y. Fang, "Webshell detection based on random forest-gradient boosting decision tree algorithm," in *Proc. 2018 IEEE 3rd International Conference on Data Science in Cyberspace*, 2018.
[7] K. Y. Wu, Z. M. Zheng, and S. T. Tang, "BVDT: A boosted vector decision tree algorithm for multi-class classification problems," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 31, no. 5, 2017.
[8] Y. Cho, J. Yoon, and S. Lee, "Using social network analysis and gradient boosting to develop a soccer win-Close prediction model," *Engineering Applications of Artificial Intelligence*, vol. 72, pp. 228-240, June 2018.
[9] X. C. L, Y. Cheng, C. Gao, and L. S. Chen, "Discovering pollution sources and propagation patterns in urban area," in *Proc. KD'17.* pp. 13-17, August 2017.
[10] Y. Zheng, F. R. Liu, and H. P. Hsieh, "U-Air: When Urban Air Quality Inference Meets Big Data," in *Proc. the 19th SIGKDD conference on Knowledge Discovery and Data Mining*, August 2013.
[11] K. Ibrahim, and O. Suat, "A deep learning model for air quality prediction in smart cities," *IEEE .International Conference on Big Data*, December 2017.
[12] L. Li, Y. H. Liu, and M. Cai, "A forecast model for urban air quality based on meteorological similarity criteria," *Environmental Science & Technology*, vol. 36, no. 5, pp. 156-161, 2013.
[13] C. L. Wu, Q. Li, and J. X. Hou, "PM2.5 concentration prediction using convolutional neural networks," *Science of Surveying and Mapping*, vol. 43, no. 239, pp. 1-12, May 2018.
[14] J. P. Keller, C. Olives, and S. Y. Kim, "A unified spatiotemporal modeling approach for predicting concentrations of multiple Air pollutants in the multi-ethnic study of atherosclerosis and air pollution," *Environmental Health Perspectives*, vol. 123, no. 4, pp. 301-309, October 2015.
[15] X. Feng, Q. Li, and Y. J. Zhu, "Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation," *Atmospheric Environment.* vol. 107, pp. 118-128, 2015.
[16] L. M. Wang, X. H. Wu, and T. L. Zhao, "A scheme for rolling statistical forecasting of PM25 concentrations based on distance correlation coefficient and support vector regression," *Acta Scientiae Circumstantiae*, vol. 37, no. 4, pp. 1268-1276, August 2017.
[17] T. K. Kim, B. Ignas, and C. Roberto, "Making a shallow network deep: Conversion of a boosting classifier into a decision tree by boolean optimisation," *International Journal of Computer Vision*, vol. 100, no. 2, pp. 203-215, 2012.
[18] J. F. Hu, and J. L. Min, "Automated detection of driver fatigue based on EEG signals using gradient boosting decision tree model," *Cognitive Neurodynamics*, vol. 12, no. 4, pp. 431-440, April 2018.
[19] H. D. Pham, D. L. Tuan, and N. V. Thanh, "Static PE malware detection using gradient boosting decision trees algorithm," in *Proc. International Conference on Future Data and Security Engineering*, pp. 228-236, November 2018.
[20] J. B. Gong, X. X. Gao, H. Cheng *et al.*, "Integrating a weighted-average method into the random walk framework to generate individual friend recommendations," *Science China*, vol. 60, November 2017.
[21] Y.Y. Zhao, B. Qin, and T. Liu, "Encoding syntactic representations with a neural network for sentiment collocation extraction," *Science China*, vol. 60, November 2017.
[22] S. Zhao, X. M. Liu, Z. Duan, Y. P. Zhang *et al*., "A survey on social ties mining," *Chinese Journal of Computers*, vol. 40, no. 3, pp. 535-554, 2017.

Jibing Gong received his PhD degree in Institute of Computing Technology, Chinese Academy of Sciences, China. He is a professor in the School of Information Science and Engineering, Yanshan University. He is head of the Knowledge Engineering Group research team in Yanshan University. His main research interests include big data analytics, heterogeneous information network, machine learning and data fusion.

He is a member of the China Computer Federation Technical Committee on Internet of Things, a member of the Language and Knowledge Computing Committee of the Chinese Information Processing Society and a member of the Social Media Processing Committee of the Chinese Information Processing Society of China. He awarded the first prize of the Chinese Society of Artificial Intelligence for scientific and technological progress. He published over 40 scientific articles in the country and abroad.

Dan Wang was born in Bao Ding, Hebei Province on 28 February 1993. She is a graduate student in the School of Information Science and Engineering, Yanshan University. She was admitted to bachelor's degree in computer science and technology in 2017. Her main research interests include machine learning and news mining.

Chen Da was born in Bao Ding, Hebei Province on 28 August 1993. He is a graduate student in the School of Information Science and Engineering, Yanshan University. He was admitted to bachelor's degree in computer science and technology in 2017. His main research interests include machine learning and big data.

Shuli Wang is a lecturer with a master of science degree. He entered Shandong University in 1998 and began his undergraduate study in information and computing science. He worked in Yanshan University after graduation in 2002 and continued his postgraduate study in probability theory and mathematical statistics in Yanshan University in 2008. The main research direction is machine learning and materials analysis.