Comparison of Predictions of Air Pollutants Concentrations in Portland and Fuzhou by Various Models

Abstract

The rapid development of low-cost sensors greatly improves the accuracy of their monitoring data. At the meanwhile, with the increasing popularity of low-cost sensors, the amount of air pollution monitoring data in the time and space dimensions is growing rapidly. In such case, the deep integration of air pollution monitoring and machine learning that rely on big data will become an inevitable trend. One of the ways is to predict the future concentrations of air pollutants by learning the monitoring data that we acquired using machine learning. In this study, We predicted four concentrations of air pollutants in Portland U.S.A. and Fuzhou China. Four concentrations of air pollutants: CO, NO_2, O_3, $PM_{2.5}$ .We used multiple methods including liner model: Liner-Regression(LR); nonlinear model: Artificial Neural Network (Long-Short-Term-Memory LSTM), Decision Tree (DT), Support Vector Regression (SVR) and integrated algorithm: Gradient Boosting Decision Tree (GBDT), Random Forest Regression(RFR). We evaluated the performances of each model by calculating coefficients of determination (R^2), mean normalized error(MNE%) and mean normalized bias(MNB%). The experimental results demonstrate that: GBDT and RFR performed well in the prediction of CO and NO_2. ANN and RFR yielded the best performance in prediction of O_3. For the prediction of $PM_{2.5}$, ANN, GBDT, RFR can get good accuracy. Then, we studied the cross interference of temperature, humidity and other pollutants on the prediction of pollutant concentrations. The experimental results demonstrate that: temperature and humidity have positive impact on the prediction of CO concentration; NO_2 has positive impact on O_3 concentration prediction; humidity has significant positive impact on $PM_{2.5}$ concentration prediction.

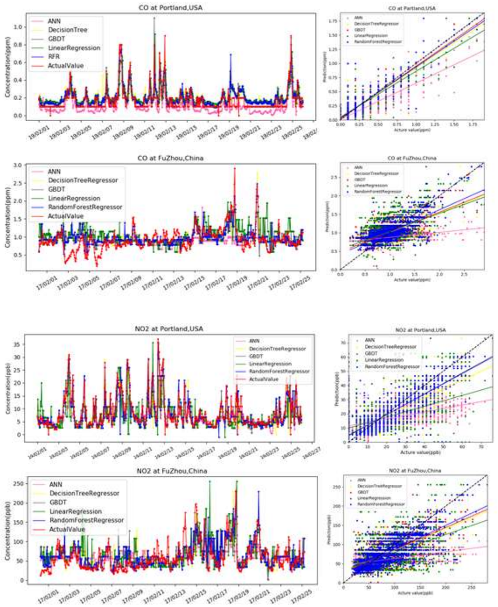
Methods

1. Normalization is required so that all the inputs are at a comparable range to solve the cross interference issue, also because the characteristics of SVR and ANN model.

2. After some comparisons, we selected 2-layer-LSTM model and DT,RFR model with maximum tree depth of 8 to capture the information in pollutant concentration series.

3. Models training at the small data, in such case, overfitting is a serious problem in such models. We limit the maximum depth of the tree and utilize Dropout and L2-regularization technique to address this problem, at the same time maintain the prediction accuracy.

4. For the above models, manually selecting witch pollutant should be input to model based on the prior knowledge will greatly improve the accuracy of prediction. Because this can minimize the fluctuations bring by unrelated parameters.



CO

GBDT(R^2=0.7) and RFR(R^2=0.7) get the better performance at CO prediction in both Portland and Fuzhou. The reason is that: the range of CO concentration is in a small interval. It is a good condition for decision tree model with a certain tree depth.

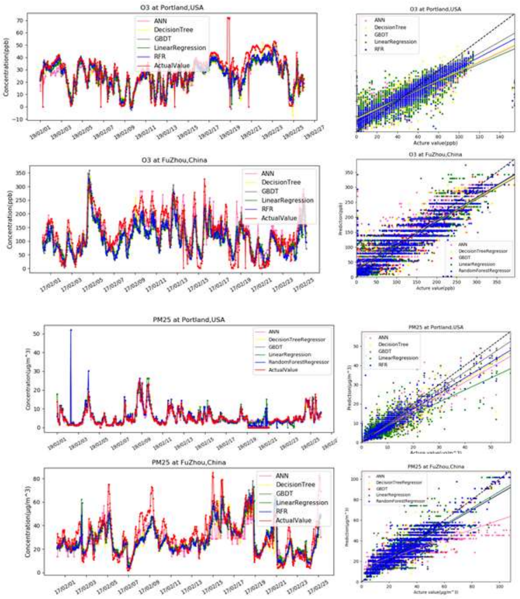
By comparison, NO_2, O_3, $PM_{2.5}$ have no positive effect on the prediction of CO.

Temperature and humidity play positive roles in the prediction of CO concentration. Max(R^2) from 0.70 to 0.75

NO_2

GBDT gets the best performance at CO prediction in both Portland and Fuzhou.

CO, NO_2, O_3, $PM_{2.5}$, temperature and humidity have no positive effect on the prediction of NO_2.



O_3

ANN and RFR get the best performance at O_3 prediction in both Portland and Fuzhou.

NO_2, humidity play positive roles in the prediction of O_3. Max(R^2) from 0.65 to 0.70

Temperature and humidity have no significant effect on CO concentration prediction. Max(R^2) from 0.70 to 0.71.

$PM_{2.5}$

All models can show good results in $PM_{2.5}$ concentration prediction. The optimal model are RFR and GBDT.

CO has a positive effect on $PM_{2.5}$ concentration prediction. Max(R^2) from 0.74 to 0.77.

O_3,NO_2 have no significant effect on $PM_{2.5}$ concentration prediction.

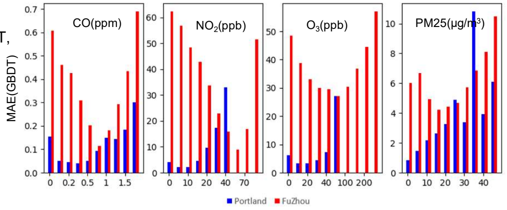
Humidity plays a significant positive role in the prediction of $PM_{2.5}$. Max(R^2) from 0.74 to 0.82.

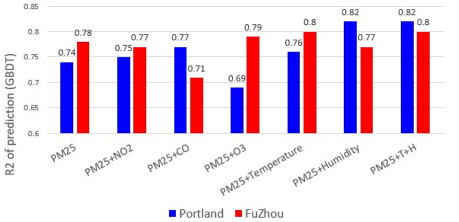
Analysis of results

Target Gas	Model Type	R^2		MNE (%)		MNB (%)	
		Portland,USA	FuZhou,China	Portland,USA	FuZhou,China	Portland,USA	FuZhou,China
CO	ANN	0.6	0.0	40.3	33.0	-29.7	10.7
	Decision Tree	0.6	0.3	36.2	29.9	20.7	9.4
	GBDT	0.7	0.2	28	30.0	14.9	12.1
	LinearRegression	0.5	-0.1	45.5	34.0	25.4	12.9
	RandomForestRegressor	0.7	0.3	32.5	29.0	20.6	10.0
	SVR	0.2	-0.2	98.4	45.3	94.8	34.4
NO_2	ANN	0.3	0.2	44.6	52.1	6.6	24.5
	Decision Tree	0.2	0.4	43.6	45.5	16.1	25.1
	GBDT	0.4	0.5	45.4	41.8	24.1	21.3
	LinearRegression	0.2	0.3	60.3	45.9	15.3	20.4
	RandomForestRegressor	0.3	0.5	38.8	39.8	16.5	19.7
	SVR	0.2	0.0	67.7	66.7	44.1	44.7
O_3	ANN	0.7	0.6	63.4	154.7	42.5	154.7
	Decision Tree	0.5	0.7	76.7	133	53.2	133
	GBDT	0.6	0.5	46.8	12.7	29.9	12.7
	LinearRegression	0.6	0.6	81.3	15.9	60.5	15.9
	RandomForestRegressor	0.7	0.7	57.1	12.5	40.1	12.5
	SVR	-0.8	0.6	111.6	16.8	108.7	16.8
$PM_{2.5}$	ANN	0.8	0.5	25.1	25.1	-3.4	-3.4
	Decision Tree	0.7	0.7	18.7	30.3	6.7	5.5
	GBDT	0.8	0.7	17.3	22.9	3.7	-1.5
	LinearRegression	0.6	0.7	27.2	34.5	-7.2	9.2
	RandomForestRegressor	0.8	0.8	19.5	24.9	7.2	3.6

For each pollutant, the best model is marked in red. In general, the integrated algorithm related to decision trees: GBDT, and RFR perform better than other models. ANN model performed better when predicting the concentration of O_3 and $PM_{2.5}$. SVR showed the worst performance in each prediction.

Right image shows the MAE values predicted by GBDT in different concentration ranges.





Left bar chart plot the R_2 of $PM_{2.5}$ prediction in Portland and Fuzhou by GBDT model. The experiment selected different input parameters. We discussed the effects of NO_2, CO, O_3, temperature and humidity on PM25 prediction. At Portland, It can be seen that humidity has the greatest impact on $PM_{2.5}$ prediction, followed by CO and temperature. O_3 has a negative impact on $PM_{2.5}$ prediction. At Fuzhou, temperature has the greatest impact on $PM_{2.5}$ prediction.