Geophysical Research Abstracts
Vol. 21, EGU2019-3121-1, 2019
EGU General Assembly 2019

Air Pollutant Prediction: Comparisons between LSTM, Light GBM and Random Forest

Peng Chen (1,2), Aichen Niu (1,2), Wei Jiang (3), and Duanyang Liu (4)

(1) Key Laboratory of Transportation Meteorology of China Meteorological Administration [U+FF0C] Nanjing [U+FF0C] China (409856986@qq.com), (2) Jiangsu Meteorological Information Center, Meteorological Bureau of Jiangsu Province, China, (3) Jiangsu Climate Center, Meteorological Bureau of Jiangsu Province, China, (4) Jiangsu Meteorological Observatory, Meteorological Bureau of Jiangsu Province, China

Concentrations of air pollutants are increasing along with societal development and have the potential to be a serious human health risk (e.g. heart and lung disease). Thus, it is necessary to predict and evaluate air pollutant concentrations and take certain actions to reduce their impacts on individuals. In this study we investigated the prediction of PM2.5 and PM10 concentrations based on three methods—namely, LSTM, Light GBM, and Random Forests.

Nanjing and Changzhou are cities located in southeastern China. As an intensively industrial city, Nanjing consumes larger amounts of coal compared to other cities with similar sized economies. In 2012, it suffered 226 days of haze, which was the most in China in that year. The high levels of coal consumption and industrial production cause excessive emissions of air pollutants. Compared to Nanjing, Changzhou is a city on a relatively smaller industrial scale. Annually, this city consumes far smaller amounts of coal and suffers fewer days of haze compared to Nanjing.

The data used in this study are hourly air pollutant concentrations and meteorological variablesfor the cities of Nanjing and Changzhou from January 2013 to May 2018. The air pollutant concentration data are from the Jiangsu Environment Monitoring Center, located in Nanjing, while the meteorological data—including temperature, dewpoint temperature, u-wind, and 2-m wind speed—are from automatic weather stations.

We applied three approaches to forecasting the air pollutant concentrations: LSTM, Light GBM, and Random Forests. The LSTM network is a special kind of RNN developed by exploding and vanishing gradients while training traditional RNN networks. Light GBM is a new, highly efficient gradient-boosting decision tree that has been widely used by several winners of various machine-learning competitions. Random Forests is an ensemble learning method for classification and regression, operated by developing several decision trees analyzing sets of variables. It is helpful in correcting the habit of decision trees to be overfitted to the training set.

Compared to LSTM and light GBM, Random Forests showed higher accuracy , and performed best with respect to the statistical parameters used to evaluate the predictions(RMSE, MAE, and R2).The results showed that the prediction of the PM2.5 concentration for both cities was the better accurate among all the four air pollutants examined.

Random Forests has several advantages compared with the other two methods, which are based on a gradient learning algorithm. Most notably, Random Forests performs best in terms of the accuracy and generalization of the output. On the other hand, since machine learning is data-driven, it requires large amounts of data to obtain a precise result. Also, to apply the algorithmic process to solve problems in other regions, the correct features of these regions will need to be learnt via large quantities of data. Nonetheless, it should be possible for Random Forests models to be further optimized with hyper-parameters to achieve even better results. In the future, we will try to use a deeper network to forecast these elements. Combining Convolutional Neural Networks and adding spatial correlation may improve the applicability and accuracy of predictions.