Research on the application of urban air quality prediction and prediction model under the background of big data

Li Liu*

Shandong Xiehe University, Yaoqiang Town, Licheng District, Jinan City, China

Abstract. In the previous research on air quality prediction, the research on the problem is usually one-sided, and many problems are solved from a single time dimension. In the research of this problem, this paper starts from the time dimension and the space dimension respectively. Considering the temporal continuity and spatial diffusion of air pollutants, the prediction results of the two dimensions are dynamically combined. Comprehensive consideration of various factors to achieve better prediction results. In order to solve the problem that there are few air quality monitoring stations in cities and there is no monitoring data in a large number of areas, an air quality prediction model is proposed.

1 Introduction

With the continuous development of urbanization, the scale of the city is expanding day by day, the number of cars in the city is increasing, and the air pollutants from factories around the city are discharged, resulting in the deterioration of the urban environment. Urban air quality has gradually become a topic of concern for urban residents. How to use historical air quality data, meteorological data, POI data, weather forecast and other urban big data to reasonably design an air quality prediction and speculation model to better help urban residents plan outdoor travel arrangements and route planning of outdoor activities is an urgent need at present.

When urban residents choose to exercise outdoors for a long time and long distance, they often need to plan routes in advance, and outdoor air quality is also an important indicator that needs attention. We can get the distribution map of urban air quality in the next 6 hours according to the starting and ending points of users' outdoor travel and in combination with the spatial distribution of air quality. With the help of the map route planning function, the weight is set for the distance and AQI value and the route is scored, so as to recommend the route with better comprehensive distance length and AQI for users.

2 Air quality prediction model

We propose an air quality prediction model based on deep learning. The air quality prediction

* Corresponding author: 9989037@163.com

model aims to solve the problem of air quality prediction from the time dimension and the space dimension through the time predictor and the space predictor. Two predicted values are obtained by comprehensively considering the diffusion law of pollutants in the two dimensions. The prediction aggregator combines the predicted values of the two predictors to dynamically integrate the results of the two predictors to further improve the prediction accuracy.

2.1 Model architecture

Air pollutants exist in time continuity and space diffusion, so we consider to construct two predictors respectively from the time dimension and the space dimension: the time predictor and the space predictor. On the one hand, because the value of air pollution changes with the passage of time, the time predictor pays more attention to the prediction of time series changes of monitoring stations. The AQI value of the station in the past few hours, the meteorological data of the area where the station is located and the meteorological forecast data are used to predict the AQI value at the future time. LSTM has a good prediction effect on sequence data, so we choose to use LSTM model to build a time predictor for regression prediction of AQI value. On the other hand, the space predictor pays more attention to the diffusion of air pollution in space. After the air pollutants are discharged into the atmosphere, they will diffuse to the surrounding areas under the influence of wind and other meteorological factors. Considering that it is difficult to extract these features manually, we use deep neural network to construct spatial predictor. Then the meteorological data and the historical data of the surrounding stations are input into the space forecaster to learn the impact of the surrounding stations on the target stations. In order to further improve the prediction accuracy, we combine the results of time predictor and space predictor. However, under different meteorological conditions, the result weights of time predictor and space predictor are different. For example, when the wind speed in a certain place is large and the pollutants diffuse rapidly, the weight of the spatial predictor will be slightly larger. Therefore, after obtaining the results of time prediction and space prediction, we put the results of the two predictions into the decision tree. The decision tree will dynamically adjust the weights of the two prediction results according to the current meteorological conditions, and integrate them to obtain the final prediction results.

2.2 Time predictor

2.2.1 Feature extraction

On the time predictor, we consider that the data used include AQI data, meteorological data and weather forecast data. Through the existing data sets, we need to extract the features that are effective for AQI prediction.

We use the keras tool to build the time predictor. The time predictor is composed of LSTM model, including three layers of neural network, one input layer, one hidden layer and one output layer. The number of nodes in the input layer is determined by the dimension of the input data, and the data output from the output layer is the AQI value predicted for 6 hours respectively.

2.2.2 Model training

Before training, we randomly divided the whole data set into training set (70%), verification set (10%) and test set (20%). During the training, AQI data of the first 3 hours are input into

the time predictor; The AQI data of the last 6 hours is used as a label to verify the output error of the time predictor. During the test, we use different eigenvectors as the input of the time predictor to get the predicted AQI value for the next 6 hours. The time predictor inputs the meteorological data, weather forecast data, AQI historical data and the time and time characteristics extracted from the time into the model, and the model will output the AQI forecast value in the next 6 hours. However, the disadvantage of time predictor is that it can not deal with the problem of air pollutant diffusion in spatial dimension.

2.3 Spatial predictor

In addition to the continuation of time, the air pollution of an area will also be affected by its surrounding areas.That is, the AQI value of the target site to be predicted is related to the AQI value of the surrounding sites. Therefore, the AQI values of surrounding stations should be used as characteristics. Surrounding stations include not only nearby stations, but also stations located in surrounding cities. The distance between the target site and the surrounding sites ranges from several kilometers to hundreds of kilometers. It is worth noting that surrounding stations with different distance and direction from the target station have different effects on the AQI of the target station. Taking the target station as the center, we draw three circles with radius of 30km, 150km and 300km respectively.Each circle is divided into eight sectors, so a total of 24 sectors are planned. Each sector corresponds to a geographical area, and air quality monitoring stations in Beijing and surrounding cities are assigned to the corresponding area. Different colors of sectors represent different ranges of AQI values. If no monitoring station is deployed in the sector, the sector is marked as transparent.

2.4 Predictive aggregator

Prediction aggregator dynamically integrates time predictor and space predictor to predict the target monitoring site. Space and time forecasters use different dimensions to predict the air quality at the same site, providing different prediction angles. In different cases, the prediction weights of time dimension and space dimension are different. For example, when the wind speed of the station to be measured is very high, the spatial diffusion will be faster, so the weight of the spatial predictor is relatively higher. Therefore, we need to choose a suitable method to dynamically integrate the results of the two predictors according to the current actual situation.

Considering that these two prediction results are relatively one-sided from their respective perspectives, we should consider these two results comprehensively. The prediction aggregator we proposed uses classification regression tree to aggregate the data of temporal prediction and spatial prediction. Because it can automatically determine the importance of features, and the amount of calculation is small. The basic idea of prediction aggregator is to build a tree by using the characteristics of time predictor and space predictor. We combine the result data of time predictor and space predictor with weather forecast data and meteorological data into a vector. Then, it prunes branches by learning samples. Finally, the optimized subtree with the smallest error is selected to output the predicted AQI value. When a piece of data is input into the prediction aggregator, the model gives different weights to the time predictor and the space predictor respectively according to the meteorological and weather forecast data of the target station, and finally gets a more accurate AQI prediction value.

3 Air quality prediction model

Through the air quality prediction model, we can get the prediction value of the future time of the area where the detection station is located, but for the area where the station is not deployed, we cannot get their air quality value. Therefore, the problem of air quality estimation needs to be solved. In the problem of air quality estimation, the traditional solution is Kriging interpolation and inverse distance weighted interpolation. However, when the number of sample points is insufficient, the general spatial difference method will have large errors and inaccurate speculation. Therefore, our solution to the problem of air quality prediction is the air quality prediction model based on the third law of Geoscience. The third law of Earth Science is a method to solve the spatial speculation proposed by Professor Zhu Axing in the United States. This method has little limitation on the number of samples needed for spatial prediction, and does not require specific spatial distribution of samples to achieve high-quality prediction.

4 Conclusion

With the rapid development of cities, the data volume of urban big data increases rapidly. How to solve the existing problems in cities through urban big data is an urgent research point. Through multi-source city big data and data mining and deep learning methods, this paper studies the prediction and prediction of air quality closely related to the hot issue of urban air quality. Through the prediction model, the predicted value of the air monitoring station area in the next 6 hours can be obtained. By calculating each grid in the city through the air quality prediction model, the distribution map of urban air quality in the next 6 hours can be obtained. With the help of map route planning function, the distance and AQI value are given weighting and the route is scored, so as to recommend a route with better comprehensive distance length and AQI for users.

2021 Shandong Provincial Statistical Application Research project: Application research of Urban Air Quality Prediction and Prediction Model in the context of Big Data (2021TJYB020)

References

1. Zheng Y, Yi X, Li M, et al. Forecasting fine-grained air quality based on big data [C]// Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015: 2267-2276.
2. Bai Y, Li Y, Wang X, et al. Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions [J]. Atmospheric pollution research, 2016, 7(3): 557-566.
3. Yeganeh B, Motlagh M S P, Rashidi Y, et al. Prediction of CO concentrations based on a hybrid Partial Least Square and Support Vector Machine model [J]. Atmospheric Environment, 2012, 55: 357-365.
4. Zhu A X, Lu G, Liu J, et al. Spatial prediction based on Third Law of Geography [J]. Annals of GIS, 2018, 24(4): 225-240.
5. Chu H-J, Bilal M. PM 2.5 mapping using integrated geographically temporally weighted regression (GTWR) and random sample consensus (RANSAC) models [J]. Environmental Science and Pollution Research, 2019, 26(2): 1902-1910.
6. Wortley R, Townsley M. Environmental criminology and crime analysis [M]. Taylor & Francis, 2016.