Application of data mining to the analysis of meteorological data for air quality prediction: A case study in Shenyang

Chang Zhao, Guojun Song
School of Environment& Natural Resources, Renmin University of China
1804097211@qq.com

Abstract. Air pollution is one of the important reasons for restricting the current economic development. PM2.5 which is a vital factor in the measurement of air pollution is defined as a kind of suspended particulate matter with its equivalent diameter less than 25μm, which may enter the alveoli and therefore make a great impact on the human body. Meteorological factors are also one of the main factors affecting the production of PM2.5, therefore, it is essential to establish the model between meteorological factors and PM2.5 for the prediction. Data mining is a promising approach to model PM2.5 change, Shenyang which is one of the most important industrial city in Northeast China with severe air pollutions is set as the case city. Meteorological data (wind direction, wind speed, temperature, humidity, rainfall, etc.) from 2013 to 2015 and PM2.5 concentration data are used for this prediction. As to the requirements of the World Health Organization (WHO), three data mining models, whereby the predictions of PM2.5 are directly generated by the meteorological data. After assessment, the random forest model is appeared to offer better prediction performance than the other two. At last, the accuracy of the generated models are analysed.

1. Introduction

Air pollution is one of the most serious problems that have an impact on the economy. In recent years, extreme air pollutions happened frequently in northeast of China. Shenyang, which is an important industrial city with large populations, suffer from serve air pollution. The concentration of PM2.5 is a vital factor in the measurement of air pollution.

PM2.5 is defined as a kind of suspended particulate matter with an equivalent diameter less than 25μm, the size of the particles determines the location of it in the human body, if the equivalent diameter of particles is less than 2.5μm, and the particles may enter the alveoli and therefore make a great impact on the human body. Particulate air pollution including PM2.5 is likely to be one of the most important factor to respiratory disease, decreased lung function and increased hospitalizations [1, 2].

There are many reasons for the production of PM2.5, in addition to industrial emissions and traffic generated pollutants, meteorological factors are also one of the main factors affecting the production of PM2.5 [3]. Therefore, it is essential to establish the model between meteorological factors and PM2.5 for the prediction.

There are two general methods usually used for air quality prediction, deterministic method which employs chemical models with theoretical meteorological emissions to stimulate the pollutant transfer and diffusion processes [4, 5]. The main courses of low accuracy always come from unreliable emission data and incomplete theoretical bases [6, 7]. Compared with complicated deterministic
methods, there is no doubt that statistical methods which depend on simple statistical modelling techniques like multiple linear regression (MLR) and neural networks is more available. Therefore, data mining is a promising approach to model PM2.5 change, it is proved to be a reliable method as numerous successful applications of data mining used for predictions in many fields are reported [8, 9]. In this study, three algorithms in data mining, logistic regression, linear discriminant analysis (LDA) and random forest are used to predict PM2.5 concentration in different air conditions.

2. Methodology

Data mining is a rapidly growing field of research of the past twenty years. A prediction model is built to determine future outcomes rather than the current ones, in this study, three models with categorical output attributes are built.

2.1. Logistic regression

Logistic regression is a kind of nonlinear regression algorithm which associates a conditional probability score with each data instance. A possibility value of 1 refers to one class while the value of 0 means the second class. The equation of a logistic regression model is written as a conditional probability,

\[ p(y = 1 | x) = a_1 x_1 + a_2 x_2 + a_3 x_3 + \cdots + a_n x_n \]

The equation above shows \( p(y = 1 | x) \) as an unbounded value denoting the conditional probability of seeing the class associated with \( y=1 \) given the values contained in feature \( x \). To eliminate the boundary problem, the probability is transformed into a ratio, specially

\[ \frac{p(y = 1 | x)}{1 - p(y = 1 | x)} \]

For any \( x \), the odds indicate how often the class associated with \( y=1 \) is seen relative to the frequency in which the class associated with \( y=0 \) is observed. That is:

\[ \ln \left( \frac{p(y = 1 | x)}{1 - p(y = 1 | x)} \right) = ax + c \]

Where \( x=(x_1, x_2, x_3 \ldots, x_n) \)

Finally,

\[ p(y = 1 | x) = \frac{e^{ax+c}}{1 + e^{ax+c}} \]

2.2. Linear discriminant analysis (LDA)

The basic principle of linear discriminant analysis is to distinguish any two categories through a decision-making boundary or hyperplane-type boundary in an n-dimensional sample space.

\[ Y = a_1 x_1 + a_2 x_2 + a_3 x_3 + \cdots + a_n x_n \]

Where: \( Y \) is the discriminant score (discriminant), \( x \) means the variable reflecting the object's characteristics, \( n \) is the contribution of each variable, which also known as the discriminant coefficient. Discriminant analysis is to make discriminant analysis through the discriminant functions. According to the number of sample space categories, according to these descriptive variables, the sample space is divided into two categories by a discriminant function, and the category of the predicted sample is determined by calculating the score of each discriminant function.

2.3. Random forest

The random forest modelling can be viewed as an enhancement of bagging which is considered as a more justified one compared with boosting. Random forest model creations combine two approaches: instance sampling and algorithm nondeterminism which is achieved by random split selections of regression tree or decision tree growing. After at least several hundred base models are built, their individual overfitting is cancelled by the aggregating process, which makes the random forest model a
highly resistant one to overfitting. Attribute sampling may be too aggressive or the small number of trees may be insufficient to compensate for the accuracy due to the randomized split selection, which makes its functions better than single tree models. All of the analysis in the study is done with R 3.3.3

3. Methodology

3.1. Study area
Shenyang which is one of the most important industrial city in Northeast China with severe air pollutions is set as the case city, the location of Shenyang is shown in figure 1(A). Based on the IT3 transition criteria developed by the World Health Organization (WHO) [10], which is shown by the Horizontal line on figure 1 (B), take the air quality data of 2013 as an example, PM2.5 annual compliance rate of Shenyang is 64.6%, therefore the air quality forecast of Shenyang is very necessary.

3.2. Data acquisition
Meteorological data (wind direction, wind speed, temperature, humidity, rainfall, etc.) from 2013 to 2015, heating data and PM2.5 emission data are used for this prediction. 80% of the data set is used to train the model, and the remaining 20% are used to test the model predictions. As to the requirements of the World Health Organization (WHO), PM2.5 concentration data are classified. Use “qua-day” refers to the PM2.5 concentration of qualified days (which means the concentration of PM2.5 is lower than 75mg/m³), while “unqua-day” refers to the unqualified ones.

Figure 1 The location and compliance situation of PM2.5 in Shenyang

3.2. Data acquisition
Meteorological data (wind direction, wind speed, temperature, humidity, rainfall, etc.) from 2013 to 2015, heating data and PM2.5 emission data are used for this prediction. 80% of the data set is used to train the model, and the remaining 20% are used to test the model predictions. As to the requirements of the World Health Organization (WHO), PM2.5 concentration data are classified. Use “qua-day” refers to the PM2.5 concentration of qualified days (which means the concentration of PM2.5 is lower than 75mg/m³), while “unqua-day” refers to the unqualified ones.
Figure 2 PM2.5 compliance distribution and correlation matrix

As shown in figure 2(A), with the wind speed, the average temperature and the average humidity change randomly, PM2.5 concentration compliance and non-compliance points are easily distinguished (which refer to red and black points in figure 2(A)), so the prediction is feasible. As can be seen from figure 2 (B), there is no high degree of correlation between the predictors, which makes no effect on modeling. Therefore, the establishment of three models is feasible.

3.3. Results
The prediction results of the three models, Logistic regression, linear discriminant analysis (LDA), and Random forest are shown in figure 3 (A), (B) and (C). The accuracy of the three methods and the selection of the best model for PM2.5 prediction will be detailed in the next section.

3.4. Discussions and Conclusions
Accuracy and Cohen’s Kappa are used to measure the pros and cons of the forecast. As shown in table 1, where Accuracy measures the correct proportions. Suppose that $y_{pi}$ means the prediction category of the $i$-th sample, while $y_i$ means the actual category of that one, then the accuracy of all the test samples is:

$$\text{accuracy} = \frac{1}{n_{\text{sample}}} \sum_{i=1}^{n_{\text{sample}}} 1(y_{pi} = y_i)$$

where $1(x)$ means indicator function, when the prediction results is actually the same with the reals ones, the accuracy rate is 1, the more the less the accuracy rate of the lower, the lower the difference between the predicted and the true ones, the lower the result of the accuracy.

Cohen’s Kappa is used to measure the degree of coincidence between the two labelling results, then the Cohen’s Kappa is:

$$\text{Kappa} = \frac{p_0 - p_e}{1 - p_e} = 1 - \frac{1 - p_0}{1 - p_e}$$

Where $p_0$ is the observed coincidence ratio and $p_e$ is the proportion of coincidence due to randomness. When the two labeling results are fully consistent, $K = 1$, the lower the difference between the predicted and the true ones, the lower of the results, even negative.
Table 1 Comparison of results of prediction.

| Model                        | Min      | Median   | Mean     | Max      |
|------------------------------|----------|----------|----------|----------|
| Logistic Regression          | 0.7018   | 0.7706   | 0.7643   | 0.7953   |
| Linear Discriminant          | 0.7000   | 0.7412   | 0.7559   | 0.8059   |
| Random Forest                | 0.7176   | 0.7529   | 0.7718   | 0.8187   |

(B) Cohen’s Kappa

| Model                        | Min      | Median   | Mean     | Max      |
|------------------------------|----------|----------|----------|----------|
| Logistic Regression          | 0.3154   | 0.4608   | 0.4543   | 0.5214   |
| Linear Discriminant          | 0.3321   | 0.4474   | 0.4474   | 0.5558   |
| Random Forest                | 0.3403   | 0.4653   | 0.4747   | 0.5907   |

The prediction results of the three models are shown in Table 1 and Figure 3(D). Although the accuracy of the three model pairs is similar, random Forest is considered the most accurate one, which perform the role of regression and integration at the same time. The numerical results presented in the study confirmed the superiority of such an approach for all pollutants considered.

References

[1] P. Rd, D. Bates, M. Raizenne, Health Effects of Particulate Air Pollution: Time for Reassessment, Environmental Health Perspectives, Vol. 103, No. 5 (May, 1995), pp. 472-480, (1995).
[2] P. Johnson, J. Graham, Fine Particulate Matter National Ambient Air Quality Standards: Public Health Impact on Populations in the Northeastern United States, 《Environmental Health Perspectives》, 2005, 113(9):1140-1147, (2005).
[3] Z. Chen-xi, W. Yun-qi, W. Yu-jie, Z. Hui-lan, Z. Bing-qing, Temporal and Spatial Distribution of PM2.5 and PM10 Pollution Status and the Correlation of Particulate Matters and Meteorological Factors During Winter and Spring in Beijing, 《Environmental Science》, 2014, 35(2):418-427, (2014).
[4] J.I. Jeong, R.J. Park, J. Woo, Y. Han, S. Yi, Source contributions to carbonaceous aerosol concentrations in Korea, ATMOS ENVIRON, 45(2011) 1116-1125.
[5] C. Coats, Fast Emissions Modeling with the Sparse Matrix Operator Kernel Emissions (SMOKE) Modeling System, 《Air & Waste Management Association Pittsburgh Pa》, 1996, 13(1):584-588, (1996).
[6] R. Vautard, P.H.J. Builjtes, P. Thunis, C. Cuvelier, M. Bedogni, B. Bessagnet, C. Honoré, N. Moussiopoulos, G. Pirovano, M. Schaap, Evaluation and intercomparison of Ozone and PM10 simulations by several chemistry transport models over four European cities within the CityDelta project, ATMOS ENVIRON, 41(2007) 173-188.
[7] R. Stern, P. Builjtes, M. Schaap, R. Timmermans, R. Vautard, A. Hodzic, M. Memmesheimer, H. Feldmann, E. Renner, R. Wolke, A model inter-comparison study focussing on episodes with elevated PM10 concentrations, ATMOS ENVIRON, 42(2008) 4567-4588.
[8] C. Leung, K. Joseph, Sports data mining: predicting results for the college football games, 《Procedia Computer Science》, 2014, 35:710-719, (2014).
[9] Kusiak, H. Zheng, Z. Song, Wind farm power prediction: a data-mining approach, WIND ENERGY, 12(2009) 275-293.
[10] WH, Air quality guidelines - global update 2005, 《Public Health & Environment》, 2014, (2014).