Computers, Materials & Continua CMC, vol.64, no.3, pp.1929-1944, 2020

A Haze Feature Extraction and Pollution Level Identification Pre-Warning Algorithm

Yongmei Zhang[1, *], Jianzhe Ma[2], Lei Hu[3], Keming Yu[4], Lihua Song[1, 5] and Huini Chen[1]

Abstract: The prediction of particles less than 2.5 micrometers in diameter (PM2.5) in fog and haze has been paid more and more attention, but the prediction accuracy of the results is not ideal. Haze prediction algorithms based on traditional numerical and statistical prediction have poor effects on nonlinear data prediction of haze. In order to improve the effects of prediction, this paper proposes a haze feature extraction and pollution level identification pre-warning algorithm based on feature selection and integrated learning. Minimum Redundancy Maximum Relevance method is used to extract low-level features of haze, and deep confidence network is utilized to extract high-level features. eXtreme Gradient Boosting algorithm is adopted to fuse low-level and high-level features, as well as predict haze. Establish PM2.5 concentration pollution grade classification index, and grade the forecast data. The expert experience knowledge is utilized to assist the optimization of the pre-warning results. The experiment results show the presented algorithm can get better prediction effects than the results of Support Vector Machine (SVM) and Back Propagation (BP) widely used at present, the accuracy has greatly improved compared with SVM and BP.

Keywords: Deep belief networks, feature extraction, PM2.5, eXtreme gradient boosting algorithm, haze pollution.

1 Introduction

Haze is caused by many factors, such as meteorological and non-meteorological factors. In recent years, the study of problems causing haze has become more and more popular [Lai and He (2017); Perez and Gramsch (2015)]. The Beijing-Tianjin-Hebei region of

[1] School of Information Science and Technology, North China University of Technology, Beijing, 100144, China.

[2] Department of Electronic & Information Engineering, The Hong Kong Polytechnic University, 00852, Hong Kong.

[3] School of Computer Information Engineering, Jiangxi Normal University, Nanchang, 330022, China

[4] Department of Mathematics, Brunel University, London, UB8 3PH, UK.

[5] Department of Computer Science, George Washington University, Washington DC, 20052, USA.

[*] Corresponding Author: Yongmei Zhang. Email: zhangym@ncut.edu.cn.

Received: 10 March 2020; Accepted: 27 April 2020.

China has been seriously polluted by PM2.5. In order to improve the detection accuracy of haze early-warning and reduce the harm to human production and life, the paper aims to study the method of haze feature extraction and pollution grade prediction in the Beijing-Tianjin-Hebei region.

With the continuous deterioration of air quality, the haze harm is more and more serious. The traditional numerical and statistical prediction methods are mainly used in haze prediction [Liu, He and Lau (2018); Ma, Shao, Xu et al. (2018)]. Haze has the characteristics of complex causes and strong nonlinearity, so it is difficult to obtain satisfactory results by using relatively simple statistical methods to predict the variation its trend [Han, Seo, Kim et al. (2019)]. Currently some researchers [Luo and Pan (2017); Yu, Wang and Bi (2018)] have applied neural networks to its analysis and prediction. However, with the increase of samples, the neural networks have displayed some defects such as complex tuning parameters, slow convergence speed, and they easily trap in the problem of local minimization [Zhang and Li (2015)].

With the development of computer technology, haze weather prediction methods such as SVM, nonlinear regression model, multiple linear regression model, wavelet transform, and limit learning machine are used in predicting haze, though the above methods also have defects.

Because the related parameters of the SVM kernel functions have a large influence on the prediction performance, SVM itself cannot optimize the parameters, so many scholars adopt the swarm intelligence algorithms to optimize the SVM parameters [Zhang, Zhang, Chen et al. (2018)]. However, most of the research results failed to achieve the desired effects due to complex coding, slower convergence speed, and local optimal defects [Sun, Shao, Mu et al. (2014)]. The choice and expression of the factors in the nonlinear regression model are only by speculation, which limits the prediction in some cases. Multiple linear regression models ignore interaction effects and nonlinear causality [Ma, Zhao and Chen (2014); Wei, Li and Jia (2018)]. When processing large-scale data, it is difficult to select and construct the wavelet basis of wavelet transform [Johnston and Kooten (2015)]. Limit learning machine lacks effective training methods and ignores spatial structured information [Jacobs, Burgess and Abbott (2018)].

At present, the methods widely used in the field of haze prediction are mainly SVM and BP. XGBoost is a parallel computing algorithm, which has the advantages such as fast operation speed, good robustness and high prediction accuracy. It can better solve the problems of over learning, low prediction efficiency, long training time and only suitable for small samples existed in the above methods [Mishra, Goyal and Upadhyay (2015)].

Since 2006 when Geoffrey Hinton, a professor at the University of Toronto in Canada and a leading scholar in the field of machine learning, proposed the idea of deep learning in his paper published in Science [Hinton (2006)], deep learning has begun to attract extensive attention in academia, and became a great upsurge in big data as well as Artificial Intelligence (AI). Humans can identify objects from the background environment quickly and accurately. Thus, the cognitive mechanism of human beings should be applied to haze recognition. Based on human cognitive mechanism, the paper deeply analyzes the significant characteristics affecting haze, establishes the haze prediction model based on XGBoost, grades the haze classification and predicts the

future haze weather, which is beneficial to the prediction and prevention of haze.

The rest of this paper is organized as follows. In Section 2, the proposed haze feature extraction and pollution level identification pre-warning algorithm is discussed in detail. The experiment results and analysis on real data are given in Section 3. Finally, some conclusions are drawn in Section 4.

2 Proposed algorithm

In this paper, a haze prediction algorithm on the basis of feature selection and integrated learning is presented. Feature selection can effectively eliminate redundant features and improve model accuracy. eXtreme Gradient Boosting (XGBoost) algorithm can availably enhance the fault-tolerance and generalization ability of the model, thus it reduces the misjudgment rate of the prediction model.

2.1 Data set selection method

This paper takes Beijing-Tianjin-Hebei region as an example, the data sets mainly come from the hourly air quality haze data and meteorology haze data of 29 stations in China national environmental monitoring center from January 2017 to January 2020, including Tianjin, Shijiazhuang, and Fengtai, Dongcheng, Xicheng, Haidian, Chaoyang districts of Beijing etc.

2.2 Data preprocessing

The paper cleans and preprocesses data including the exception value processing and missing data filling. The existence of outliers will interfere with the model training and lead to inaccurate prediction output. The paper divides data into numerical, nominal and ordinal data. For nominal and ordinal data, due to their fixed value ranges and bounds, outlier detection only needs to determine whether the value is within a reasonable range. For numerical data, because there is no fixed value range, and some values (such as PM2.5) are not evenly distributed, and PM2.5 can be considered as abnormal values when haze weather occurs, the paper utilizes the 3δ principle of normal distribution to detect extreme abnormal values of numerical data.

Due to the inability to obtain or omit, as well as the abnormal value processing and so on, incomplete value of a certain attribute will lead to the loss of useful information in modeling, and the null value will also cause unreliable output in modeling, it is very important to fill the missing value of the data. By analyzing the data, the number of attributes with missing values, as well as the missing number and rate of each attribute are obtained. Delete the attributes with the missing rate greater than 50%, then view the distribution of the missing data for the attribute with the missing rate greater than 30%. If the missing data is relatively dense, delete the same period of training data. For the data with smaller missing rate or the attributes with larger missing rate but scattered distribution of missing data, if the attributes are numerical, this paper uses the adjacent mean filling processing, if the attributes are not numerical, the paper adopts modes to fill.

2.3 Fusion method of low-level and high-level features

At present, haze feature extraction methods mostly used are not sufficient, mainly

considering the influence of meteorological factors [Huang, Yan and Zhang (2018); Lin, Fu and Jiang (2013)]. Haze prediction is not only a meteorological cause, but also a complex relationship with other non-meteorological factors. In addition, the relationship between various haze factors of the environment is considered to be insufficient, that is, the high-level characteristics of factors affecting haze have not been taken into account by now.

In this paper, Minimum Redundancy Maximum Relevance (mRMR) algorithm is adopted to extract haze low-level features, and Deep Belief Networks (DBN) is used to extract haze high-level features. By XGBoost algorithm, the two-layer characteristics are fused, and the difficulty of determining parameters in the traditional mathematical operation fusion is avoided. Because the multi-layer features have better data expression, the prediction results are much more excellent. The experiment results also show the extracted features can improve the accuracy of prediction.

2.3.1 mRMR-based haze low-level feature selection method

Haze weather information involves multiple factors. Redundant factors can not only waste computer storage space, but also interfere with prediction accuracy. In this paper, mRMR algorithm is used to eliminate noise haze properties and obtain key haze influence elements. mRMR algorithm is a typical filtering feature selection method based on spatial search [Irina, Luca and James (2015); You, Shu and Chen (2017)]. Maximum correlation refers to the maximum correlation between features and classification variables, and minimum redundancy means the minimum correlation between features. mRMR algorithm uses mutual information, information difference and information entropy as a feature subset search strategy. The definitions of maximum correlation and minimum redundancy are shown in Eqs. (1) and (2) respectively.

$$\max D(F,c),\ D=\frac{1}{|F|}\sum_{f_i\in F} I(f_i;c) \tag{1}$$

$$\min R(F),\ R=\frac{1}{|F|^2}\sum_{f_i,f_j\in F} I(f_i;f_j) \tag{2}$$

where, F is the feature set, c is the sample category, $I(f_i; c)$ indicates the mutual information between feature f_i and category c, $I(f_i; f_j)$ represents the mutual information between features f_i and f_j. The mutual information is defined as

$$I(x;y)=\iint p(x,y)\log\frac{p(x,y)}{p(x)\,p(y)}\,dxdy \tag{3}$$

where $p(x)$ and $p(y)$ are probability densities of two random variables x and y, $p(x,y)$ is joint probability density of x and y.

In this paper, mRMR algorithm is adopted in feature selection, mainly to solve the problem that the best m features can be obtained by maximizing the correlation between features and target variables, but the best prediction accuracy may not be obtained.

General feature selection methods are basically based on strong correlation features of the target variables, but these features may also contain some redundant features. mRMR algorithm not only ensures the maximum correlation, but also removes the redundant features. The features obtained by mRMR algorithm are obviously different, and the correlation between the target variables is also very strong.

After sorting the features through mRMR algorithm, the importance and corresponding score of each feature are obtained. The importance scores of the 2nd feature (autumn), the 16th feature (time), the 17th feature (PM2.5_last), the 19th feature (NO_2), and the 10th feature (holiday) are higher. The higher the score is, the more important the feature is. Sort the scores in descending order, extract and obtain the first 16 features as the low-level features in Tab. 1.

Table 1: Feature extraction results based on mRMR algorithm

2	16	17	19	10	18	14	11
autumn	time	PM2.5_last	NO_2	Holiday	PM10	work_first	work
20	12	13	28	15	30	32	29
CO	rest_first	rest_last	fog	week_last	rain	snow	dust

2.3.2 High-level feature extraction method of haze based on DBN

Deep learning enables a broader application of machine learning as well as AI, and becomes a new research hotspot in many fields. In view of the shallow network cannot dig the original data characteristics, and monolayer feature without hierarchical structure obtained via shallow network model learning, which affects the accuracy of the predicted data.

DBN introduces the concept of back propagation, compares the obtained output with the input, and reduces the error value to zero through the back propagation mechanism. DBN is a probability generation model with multiple hidden layers. By training the network and adjusting the weights between the neural elements, the whole network can restore the input data with a maximum probability.

The essence of DBN is the process of feature learning, that is, how to get better feature expression. In this paper, DBN is used to extract high-level features of haze, and DBN can learn more high-level abstract features of haze. Fig. 1 shows the flow chart of high-level feature extraction method.

The paper adopts the DBN model composed of 3-layer Restricted Boltzmann Machines (RBM), the first layer is haze data input layer, and the last layer is the high-level feature output layer after DBN training.

In the first layer, the number of nodes corresponding to the pre-processed haze data should be set. The number of haze data is 46, so the number of nodes in the first layer is set to be 46. The number of nodes in the middle layers has great influence on the network performance. If it is too small, the correlation between the data cannot be fully mined. If it is too large, the training time of the network can be prolonged, resulting in the network performance decline. In this paper, the number of middle layer nodes is set to be 100 through a series of experiments. In order to achieve better prediction effects, the number of high-level features is consistent with the number of haze data. The node number in the last layer of DBN is set to be 46, and the final network structure is set to be 46-100-100-46. Each RBM learning rate is set to be 0.1, and the training iteration is taken as 70.

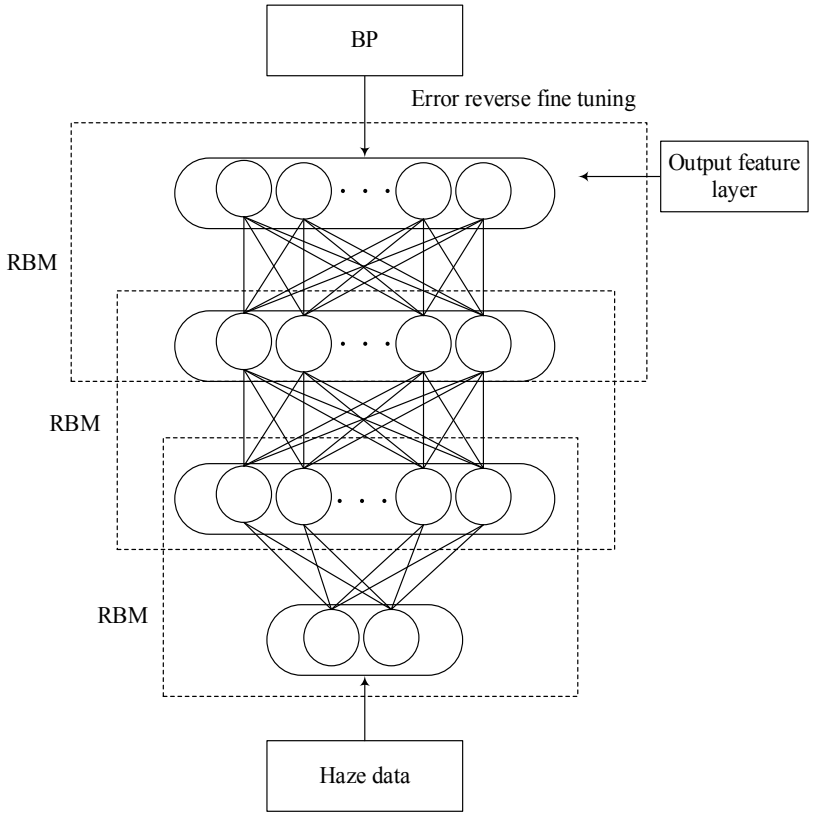

Figure 1: The flow chart of high-level feature extraction method

2.3.3 Fusion method of low-level and high-level features

Deep learning has entered a bottleneck period and the simulation of human neural structure will be the breakthrough. Geoffrey Hinton believes the key solution to overcome the limitations of AI is to build "a bridge connecting the computer science and the biology". Thus, the back-propagation is regarded as a biologically inspired breakthrough in computer science. The idea of back-propagation originates from psychology rather than engineering. Geoffrey Hinton is trying to emulate this model.

Professor Herbert Alexander Simon, one of the pioneers of AI, has been developing AI programs of cognitive psychology to reveal the essence of human cognition. He developed several programs such as LG, EPAM, GPS, proposed the famous "physical symbol system hypothesis", and finally concluded his own information-processing-oriented cognitive psychology. The great achievements in using computer to simulate human cognition led to the emergence of AI as a discipline, which made remarkable contributions to the combination of cognitive psychology and computer.

Humans can quickly and accurately identify objects from the background environment due to the effective knowledge reasoning ability and perfect vision mechanism, which are key factors for humans in complex target recognition. Therefore, human cognitive mechanism should be applied to haze recognition, the paper utilizes the human cognition

mechanism to improve the accuracy of haze prediction and pollution level identification pre-warning algorithm.

There is a huge difference between the visual features of images understood by computers and humans. Humans always recognize images on the basis of some high-level concepts which show the human's interpretation of the targets, events, and emotions expressed by images. And the interpretation represents the semantic features of images. As computer vision and AI are not perfect enough at the moment, the huge gap between the understanding of image contents by computers and that by human beings resulted in the problem of "semantic gap".

The fundamental obstacle is the "semantic gap" between low-level features and high-level semantics. The solution to the problem of "semantic gap" needs to construct the mapping and support among features at different levels through vertical correlation which enables the computers to obtain the semantic information of images accurately on the basis of low-level visual features.

Deep learning is superior to other algorithms because it can discover the distributed feature of the data due to its unique hierarchical structure and its ability to combine low-level features to form more abstract high-level features, thus making classification or prediction much easier. This paper applies deep learning in extraction hierarchical features from low-level to high-level features, which will be beneficial to solving the problem of "semantic gap" between low-level features and high-level semantics.

Although DBN can extract high-level abstract features, the more abstract features easily lose more detailed information. If only low-level features are considered, the abstract ability is insufficient, and the final prediction results are not accurate. Multi-layer feature fusion combines a variety of information to supplement different information and make prediction accuracy higher.

The main feature fusion ways include weighted sum, cascade, and so on. In this paper, 16 features extracted based on mRMR are combined with 46 features extracted by DBN, and 62 features are obtained as the input of the prediction model.

2.4 Haze prediction and pollution level identification pre-warning algorithm

XGBoost was proposed by Chen at the university of Washington in 2015. XGBoost algorithm can minimize the loss function, automatically utilize the multithreading of CPU for distributed learning and multi-core computing, and improve the computation efficiency while ensuring classification accuracy. Compared with other prediction algorithms, XGBoost has the advantages of fast speed, good effect and large-scale data processing. XGBoost has improved the loss cost function by introducing the first and second derivatives to make the model prediction results more accurate. Therefore, XGBoost algorithm is selected as the prediction model in this paper.

Haze change is a complex nonlinear system with nonlinear and abrupt characteristics [Shi, Zhu, Xia et al. (2016)]. Although DBN can predict data, the depth and complexity of the neural network are bound to affect the calculation efficiency and prediction accuracy. In order to reduce the training time of neural network and make up for the deficiencies in neural network, this paper proposes a haze prediction algorithm based on

DBN and XGBoost. XGBoost adopts extreme gradient ascent framework, it can deal with nonlinear data quickly, accurately and efficiently.

The paper presents a haze prediction and pollution level identification pre-warning model shown in Fig. 2. Obtain haze data, preprocess the data, and extract low-level and high-level features. Train the DBN model, connect the last high-level features of the DBN model with the low-level features extracted by mRMR algorithm to the XGBoost prediction model, and train the XGBoost prediction model. After the training of XGBoost prediction model is completed, the haze pollution level index is constructed, and the grade classification is on the basis of the predicted value. Finally, expert experience knowledge is added for auxiliary optimization and the final warning information is output. The specific steps for the proposed algorithm are given below.

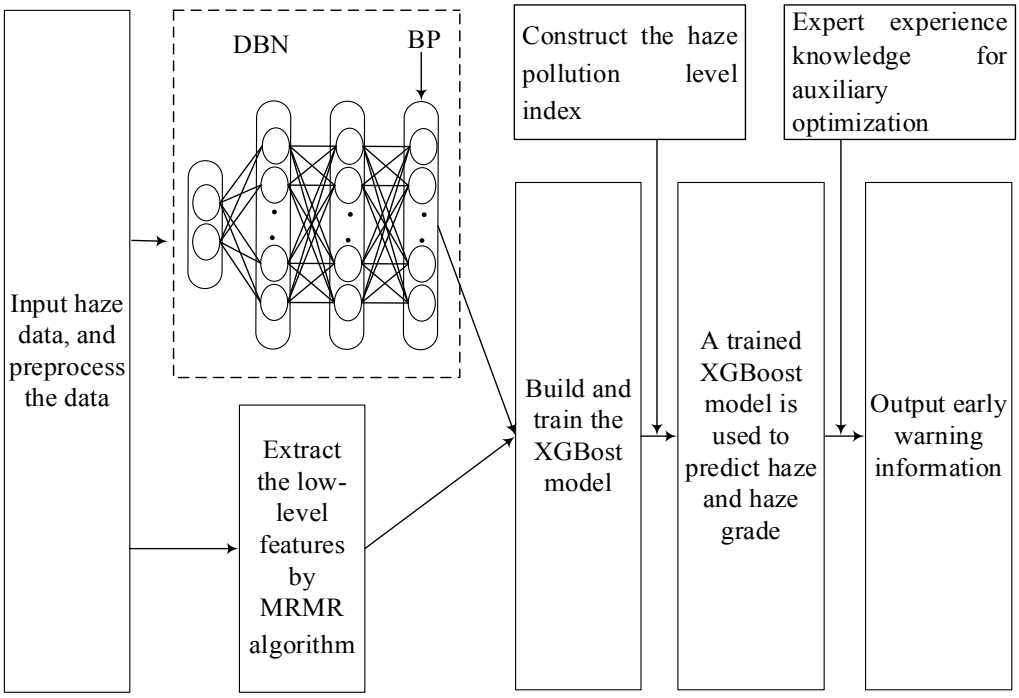

Figure 2: The haze prediction and pollution level identification pre-warning model

(1) Input haze data, clean and preprocess the data including abnormal value processing and missing value filling.

(2) Extract haze low-level features by mRMR algorithm, train DBN to extract haze high-level features.

(3) Fuse low-level and high-level features, predict haze through XGBoost algorithm. Take the low-level and high-level features as the input of XGBoost model, train XGBoost model, gradually adjust the parameters of XGBoost model by comparing the output results to optimize the model in the course of training, obtain the optimal model parameters and XGBoost model with haze prediction function. Through a lot of experiments, set the main parameters for optimal XGBoost model as shown in Tab. 2.

(4) Input test data, forecast haze on the test sets through the trained model, and output the prediction results.

(5) Construct haze pollution level indexes, grade haze level on the basis of predicted value, add expert experience knowledge to optimize haze pre-warning results, and output pre-warning information in accordance with the warning indexes.

Table 2: Main parameters for optimal XGBoost model

Parameters	eta	Nthread	subsample	Nrounds	Colsample_ bytree
Values	0.15	6	0.6	5000	0.7
Parameters	Early.stop.round	Seed	Max_depth	Booster	Objective
Values	200	42	6	gbtree	Reg:linear

In Tab. 2, the parameter eta is the learning rate, Nthread is the number of threads that can operate in parallel. The parameter subsample is to control the proportion of random sampling for each tree, Nrounds is the number of trees. Colsample_bytree is used to control the proportion of each random sampling feature for each tree, Early.stop.round is early stopping times. Seed is the seed of random numbers, and it can reproduce the results of random data. Max_depth is the maximum depth of the tree. booster is to select a base classifier, this paper sets it to be gbtree, namely a tree-based model. Objective is to define the objective function, it is set to be reg:linear in the paper, that is linear regression.

This paper reacts the severity and emergency degree of haze by pre-warning information, reminds relevant departments and the public to take corresponding measures or emergency plans and protect the safety of life and property. According to the new air quality standard of PM2.5 inspection network, the 24-hour average and standard value distribution of PM2.5, namely PM2.5 grade level is shown in Tab. 3.

Table 3: Grade level of PM2.5

Daily average concentration of PM2.5 (ug/m^3)	0-34	35-74	75-114	115-149	150-249	250-500
Air quality grade	First grade	Second grade	Third grade	Fourth grade	Fifth grade	Sixth grade
Pollution class	Excellent	Good	Mild pollution	Moderate pollution	Heavy pollution	Serious pollution

As it can be seen in Tab. 3, haze pollution levels are divided into excellent, good, mild, moderate, heavy and serious pollution. The corresponding PM2.5 values are respectively 0-34, 35-74, 75-114, 115-149, 150-249 and 250-500. To realize the level of haze pollution and the output of pre-warning, this paper needs to predict the value of PM2.5 in the haze weather, judge the value of PM2.5, give the corresponding concentration interval, output the appropriate pre-warning information, and represent the air quality level.

Deep learning based on big data has achieved unexpected achievements, but the

inexplicability of the algorithm has led to many security risks. Combining knowledge-driven and data-driven, an interpretable and robust AI theory will be established, and credible, safe and reliable AI technology will be developed, and it will effectively expand the application range of AI.

At present, deep learning can realize end-to-end learning, but the development of human civilization has accumulated rich prior knowledge so far. How to make full use of the existing human knowledge base and integrate it into the existing deep learning will become an important research focus. It is difficult to reach higher pre-warning accuracy by relying on quantitative models and current prediction analysis methods. If utilizing the expert knowledge, combining quantitative and qualitative analysis, and the pre-warning accuracy can be effectively improved.

This paper summarizes the experience and knowledge of experts on haze prediction, analyzes the historical data on haze, and obtains three rules of expert experience and knowledge, namely if (wind speed>4.5 and wind speed<7) then the haze level is excellent, if (weather=rain and humidity>94) then the haze grade is excellent, if (weather=dust and wind speed<2) then the haze level is serious pollution.

In this paper, the experience knowledge rules of experts are applied to the prediction model in auxiliary pre-warning. The prediction model is used for prediction, and the grades are divided after obtaining the specific prediction data. The expert experience knowledge is used for grade reasoning, and the results of the prediction model are corrected to obtain the final pre-warning output.

3 Experiment results and analysis

Experiments select the extracted high-level and low-level features as the input of prediction model, the hourly prediction is based on the previous hour haze features. In order to verify the performance of the proposed algorithm, the paper compares the prediction values of the model training for the haze prediction algorithm based on DBN and XGBoost, DBN, BP and SVM methods.

Taking an hour of Fengtai District as an example, Fig. 3 compares the prediction effect of the presented algorithm with the actual observation. It can be seen from Fig. 3, the prediction results are closer to the actual observation values. Although there are certain errors between the prediction and actual observation values of PM2.5, the overall curve is in higher fit degree. The prediction curve is very sensitive to the reaction of the trend when PM2.5 fluctuates greatly, and the prediction accuracy is higher.

The proposed algorithm is compared with DBN, BP, SVM, the prediction comparison effects and actual observation results are respectively shown in Figs. 4, 5 and 6. Fig. 4 shows the comparison effect of DBN directly used for prediction in Fengtai District, it can be seen the fitting degree of predicted and actual observed values is lower, and there are significant errors at most of sample points. Fig. 5 is the comparison effect of BP in Fengtai District, it can be seen the predicted values obtained by classical BP network prediction model are larger than the actual observation values, and the prediction effect is not ideal. Fig. 6 shows the comparison effect of SVM in Fengtai District, it shows the accuracy of SVM is lower, and significant errors exist in most sample points. Compared with DBN, BP and SVM, the proposed algorithm has achieved better prediction effect.

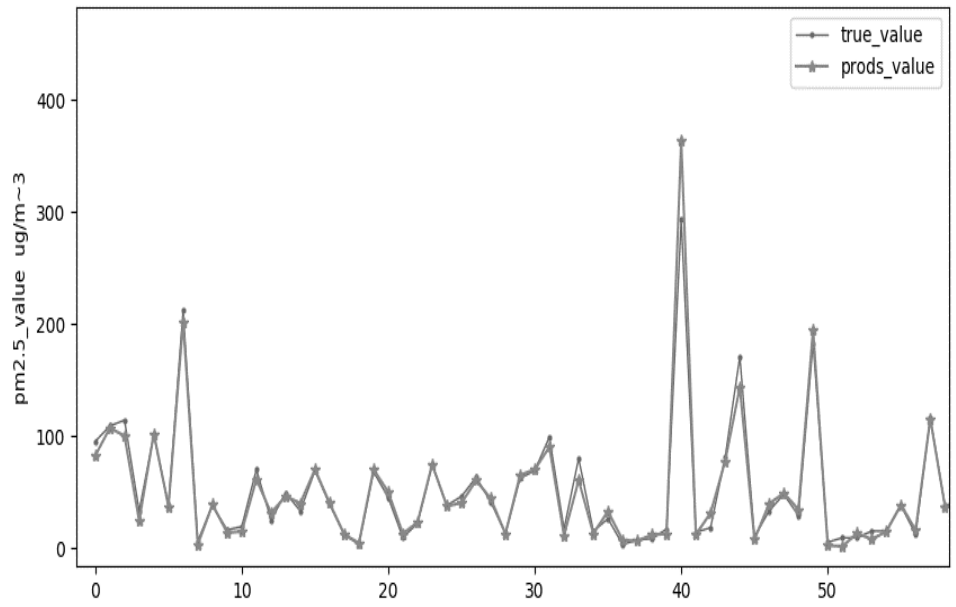

Figure 3: Prediction comparison effect of the proposed algorithm

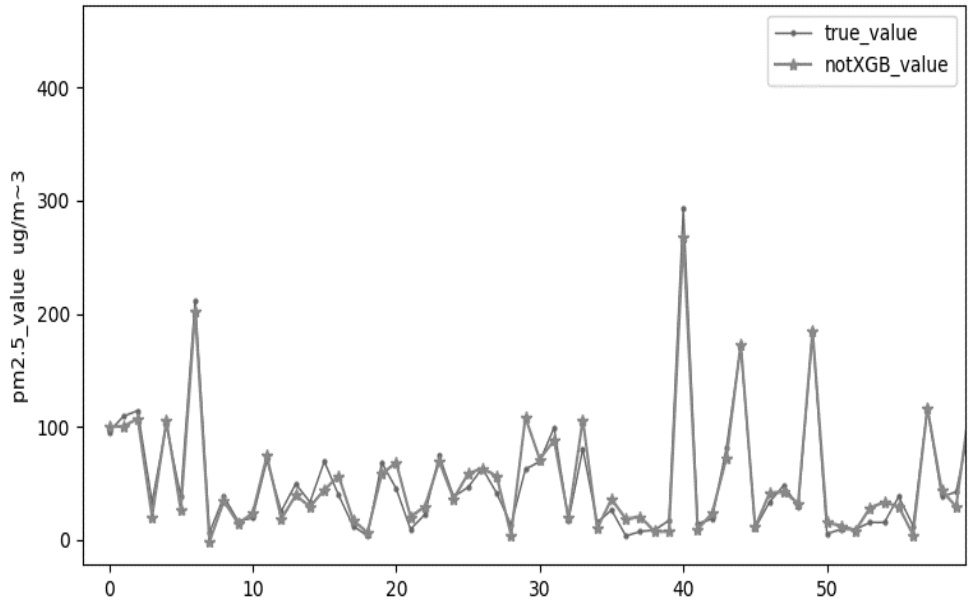

Figure 4: Prediction comparison effect of DBN

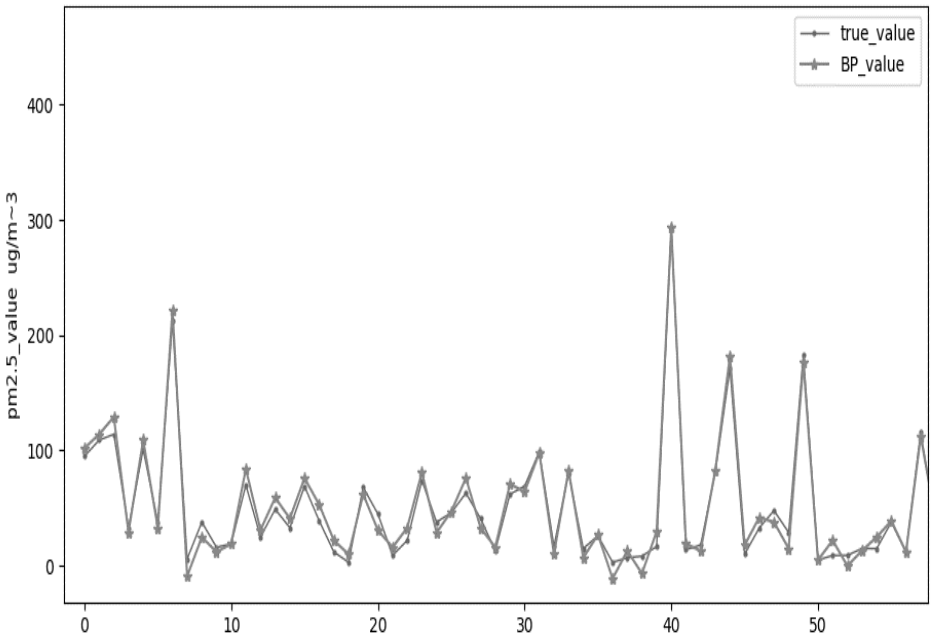

Figure 5: Prediction comparison effect of BP

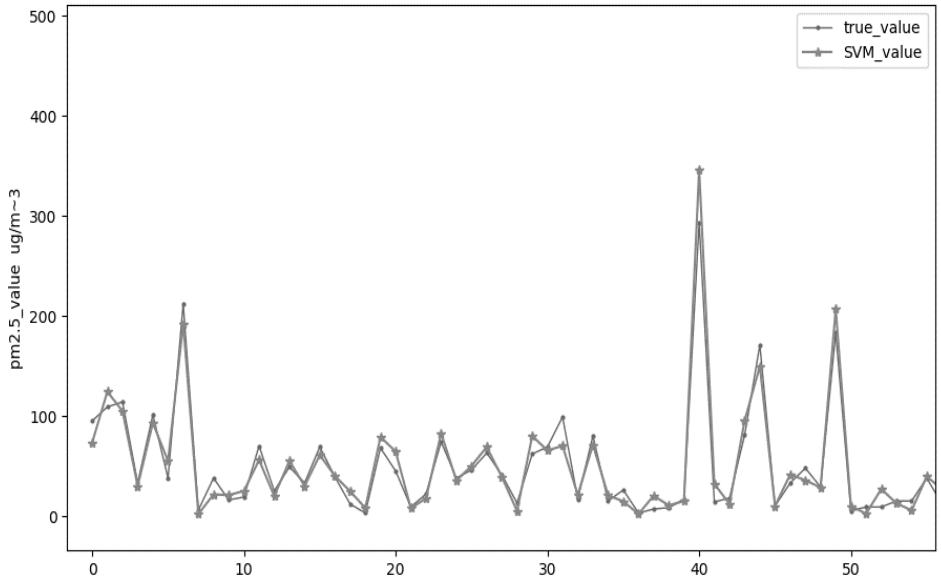

Figure 6: Prediction comparison effect of SVM

DBN prediction method only extracts the high-level features of haze. BP and SVM just extracts the low-level features of haze without fully considering the various influencing factors of haze environment and mining the correlation between the influencing factors.

DBN, BP and SVM prediction methods ignore the relationship between low-level and high-level features, resulting in poor generalization and some limitations of the prediction methods. In this paper, meteorological and non-meteorological factors are thoroughly considered to deeply mine the characteristics affecting haze. The low-level features extracted by mRMR algorithm are combined with the high-level features extracted by DBN, the low-level and high-level features are used as the input of XGBoost prediction model.

Tab. 4 shows the performance comparison of different prediction methods. This paper adopts Mean Absolute Error (MAE), Mean Square Error (MSE) and Root Mean Square Error (RMSE) as prediction error analysis indexes.

MAE represents the average value of the errors between the actual result of each data and the predicted results [Zhang, Jin, Wu et al. (2018)]. MAE is a linear fraction, the average value of the offset is directly taken, and the error of the predicted and actual values is described. It is given by Eq. (4).

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |(f_i - y_i)| \tag{4}$$

MSE shows the average value of each data error square [Zhou and Zhou (2017)], which is used to evaluate the degree of data variation, the smaller value of the MSE is better [Yin, Yuan and Zhang (2017)], as shown below:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N} (f_i - y_i)^2 \tag{5}$$

RMSE represents the average value of the square for each data error, and then calculates its arithmetic square root. RMSE is more able to punish the higher difference than MAE. RMSE is defined as

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (f_i - y_i)^2} \tag{6}$$

Table 4: Performance comparison of different prediction methods

Prediction methods	MAE	MSE	RMSE
The proposed algorithm	8.218	272.976	16.522
DBN	9.904	391.762	19.793
BP	12.836	657.819	25.648
SVM	11.710	442.345	21.032

The MAE, MSE and RMSE values of the proposed algorithm are smaller than those of DBN, BP and SVM, the error is the smallest, and the prediction accuracy is improved obviously. The results indicate the proposed algorithm outperforms the other three methods.

In addition, pre-warning accuracy comparison is shown in Tab. 5 after adding expert experience and knowledge to optimize the prediction results. The prediction accuracy without expert experience and knowledge is 83%, and the pre-warning accuracy is 88%. After adding expert experience and knowledge, the pre-warning accuracy will rise to 91%, and the pre-warning accuracy is significantly increased.

Table 5: Pre-warning accuracy comparison

The proposed prediction algorithm	Forecast accuracy	Pre-warning accuracy
Without expert experience and knowledge	0.83	0.88
Add expert experience and knowledge	--------	0.91

4 Conclusion and future work

The paper extracts low-level features based on mRMR algorithm, extracts high-level features by DBN, and effectively utilizes the complementarity of features at different levels. A haze prediction algorithm combining DBN and XGBoost is proposed. Experiment results demonstrate the algorithm can improve the prediction accuracy, and enhance pre-warning precision by adding expert experience and knowledge to optimize pre-warning information. In the future research, the characteristics of influencing haze factors will be added dynamically, and the features of small influencing factors will be replaced in time so as to improve the accuracy of haze prediction.

Funding Statement: The work was financially supported by National Natural Science Fund of China, specific grant numbers were 61371143 and 61662033, initials of authors who received the grants were respectively Z. YM, H. L, and the URLs to sponsors' websites was http://www.nsfc.gov.cn/. **This paper was supported by National Natural Science Fund of China (Grant Nos. 61371143, 61662033).**

Conflicts of Interest: The authors declare that we have no conflicts of interest to report regarding the present study.

References

Han, S. W.; Seo, J. Y.; Kim, D. Y.; Kim, S. H.; Lee, H. M. (2019): Development of cloud-based air pollution information system using visualization. *Computers, Materials & Continua*, vol. 59, no. 3, pp. 697-711.

Hinton, G. E. (2006): Reducing the dimensionality of data with neural networks. *Science*, vol. 313, no. 5786, pp. 504-507.

Huang, Y. Y.; Yan, Q. W.; Zhang, C. R. (2018): Study on the spatial-temporal change characteristics and influence factors of fog and haze pollution based on GAM. *Neural Computing and Applications*, vol. 31, no. 5, pp. 1619-1631.

Irina, D.; Luca, D. M.; James, M. (2015): PM 2.5 analog forecast and Kalman filter post-processing for the community multiscale air quality (CMAQ) model. *Atmospheric Environment*, vol. 119, no. 10, pp. 431-442.

Jacobs, E. T.; Burgess, J. L.; Abbott, M. B. (2018): The donora smog revisited: 70 years after the event that inspired the clean air act. *American Journal of Public Health*, vol. 108, no. S2, pp. S85-S88.

Johnston, C. M. T.; Kooten G. (2015): Back to the past: burning wood to save the globe. *Ecological Economics*, vol. 11, no. 120, pp. 185-193.

Lai, X. F.; He, X. H. (2017): Method based on minimum redundancy and maximum separability for feature selection. *Computer Engineering and Applications*, vol. 53, no. 12, pp. 70-75.

Lin, G; Fu, J. Y.; Jiang, D. (2013): Spatio-temporal variation of PM2.5 concentrations and their relationship with geographic and socioeconomic factors in China. *International Journal of Environmental Research and Public Health*, vol. 11, no. 1, pp. 173-186.

Liu, T.; He, G.; Lau, A. (2018): Avoidance behavior against air pollution: evidence from online search indices for anti-pm2.5 masks and air filters in Chinese cities. *Environmental Economics and Policy Studies*, vol. 20, no. 2, pp. 1-39.

Luo, J. H.; Pan, R. (2017): Research and implementation of intelligent risk recognition model based on engineering construction of neural network. *Advances in Intelligent Systems and Computing*, vol. 613, no. 8, pp. 319-326.

Ma, N.; Zhao, C. S.; Chen, J. (2014): A novel method for distinguishing fog and haze based on PM2.5, visibility, and relative humidity. *Science China*, vol. 57, no. 9, pp. 2156-2164.

Ma, X.; Shao, L. M.; Xu, G. L.; Guo, C. (2018). Foggy weather recognition based on K-means clustering algorithm. *Ship Electronic Engineering*, vol. 38, no. 12, pp. 129-133.

Mishra, D.; Goyal, P; Upadhyay, A. (2015): Artificial intelligence-based approach to forecast PM2.5 during haze episodes: a case study of Delhi, India. *Atmospheric Environment*, vol. 102, no. 2, pp. 239-248.

Perez, P.; Gramsch, E. (2015): Forecasting hourly PM2.5 in Santiago de Chile with emphasis on night episodes. *Atmospheric Environment*, vol. 124, no. 11, pp. 22-27.

Shi, Z. H.; Zhu, M. M.; Xia, Z.; Zhao, M. H. (2016): Fast single-image dehazing method based on luminance dark prior. *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 31, no. 2, pp. 1754-1768.

Sun, Z. Q.; Shao, L. Y.; Mu, Y. J.; Hu, Y. (2014): Oxidative capacities of size-segregated haze particles in a residential area of Beijing. *Journal of Environmental Sciences*, vol. 26, no. 1, pp. 167-174.

Wei, Q. Z.; Li, S.; Jia, Q. (2018): Pollution characteristics and health risk assessment of heavy metals in PM2.5 in Lanzhou. *Chinese Journal of Preventive Medicine*, vol. 52, no. 6, pp. 601-607.

Yin, Y. F.; Yuan, H. L.; Zhang, B. L. (2017): Dynamic behavioral assessment model based on Hebb learning rule. *Neural Computing and Applications*, vol. 28, no. 1, pp. 245-257.

You, M. L.; Shu, C. M.; Chen, W. T. (2017): Analysis of cardinal grey relational grade and grey entropy on achievement of air pollution reduction by evaluating air quality trend in Japan. *Journal of Cleaner Production*, vol. 142, no. 10, pp. 3883-3889.

Yu, Z.; Wang, X. C.; Bi, X. J. (2018): A light dual-task neural network for haze removal. *IEEE Signal Processing Letters*, vol. 25, no. 8, pp. 1231-1235.

Zhang, H.; Jin, X. T.; Wu, Q. M. J.; Wang, Y. N.; He, Z. D. et al. (2018): Automatic visual detection system of railway surface defects with curvature filter and improved Gaussian mixture model. *IEEE Transactions on Instrumentation and Measurement*, vol. 67, no. 7, pp. 1-16.

Zhang, Y. X.; Zhang, X.; Chen, J. Z.; Yang, J. H. (2018): Electro-mechanical impedance-based position identification of bolt loosening using LibSVM. *Intelligent Automation and Soft Computing*, vol. 24, no. 1, pp. 81-87.

Zhang, Y.; Li, Z. Q. (2015): Remote sensing of atmospheric fine particulate matter PM2.5 mass concentration near the ground from satellite observation. *Remote Sensing of Environment*, vol. 160, no. 2, pp. 252-262.

Zhou, Y.; Zhou, N. T. (2017): The analysis of causes of haze in Beijing based on big data and prediction of PM 2.5 concentration in 2017. *Technology and Innovation Management*, vol. 38, no. 6, pp. 573-581.