Air quality prediction method based on improved wavelet denoising and LSTM

Teng Pu1*, Fang lin2, Yuting Zhao2, Zhenxiang Fu2

College of Science, Southwest University of science and technology, Mianyang 621000, China
1752560404@qq.com

Abstract: in order to optimize the traditional wavelet de-noising algorithm threshold function in continuity and noise reduction effect, an improved soft hard threshold compromise method and deep learning model are proposed to analyze the prediction performance of different deep learning models and improved wavelet algorithm. RNN, GRU and LSTM neural networks are constructed to predict the original data, the traditional wavelet denoising data and the improved wavelet denoising data. The air quality data of Chengdu were used for simulation experiment. The results of quantitative analysis showed that: for AQI, PM2.5 and O3 Oh denoising will cause the loss of useful signals. The improved wavelet threshold denoising for PM10, SO2, Co, NO2 data can improve the prediction effect of the model. The results show that LSTM model has good applicability in air quality prediction, the absolute percent error is 5.867%, and the mean square error is 4.870. The prediction performance of LSTM model is about 10%, the absolute percent error is 5.176%, and the mean square error is 5.314. The traditional threshold denoising will affect the prediction results of the model for the actual data due to the loss of high-frequency signal and pseudo Gibbs phenomenon. The results show that the improved soft hard threshold compromise method can better deal with air quality data, and the LSTM model based on improved wavelet denoising has strong applicability for air quality prediction.

1. Introduction
With the rapid development of industrialization and urbanization, as well as the substantial increase of vehicle exhaust emissions, ambient air quality, especially urban air quality, has been polluted to varying degrees, which has become a major challenge to human health. Air quality forecast has been published in China since 2001. The forecast models include mechanism model and non mechanism model, The mechanism models mainly include "urban air pollution forecast (EMH)" and "urban air pollution forecast (CAPPS)". The non mechanism model obtains the change rule of pollutant concentration by analyzing the collected historical data.Since the above models can only extract the global laws of time series, and reflect the impact of unexpected events weakly, wavelet analysis as an excellent means to capture abnormal data is applied in this paper. In practical application, when wavelet threshold denoising is applied, hard threshold function or soft threshold function are usually selected to process high-frequency signal, but the effect is not ideal because the selection of threshold function will affect the denoising effect. In this paper, the improved soft hard threshold compromise threshold function is applied to denoise air quality data, and different denoising functions and depth studies are compared and analyzed The results show that LSTM model is better than other models in mean square error and average absolute percent error. Wavelet algorithm using improved threshold has better effect than traditional algorithm.
2. Improved wavelet threshold denoising and LSTM model principle

2.1 long and short term memory neural network (LSTM)
As a deep learning algorithm, long-term and long-term memory network (LSTM) is mainly used to learn long-term dependent information of time series. The key of LSTM neural network is the flexible use of its internal control gate. An LSTM unit has three gate structures: input gates, forget gates and output gates. Through the gates structure, the information can be selectively passed through, so as to update or retain the historical information and update the cell state.

![Figure 1 Structure of LSTM neurons](https://example.com/lstm.png)

2.2 improved wavelet threshold denoising
Wavelet analysis is to obtain the low-frequency and high-frequency information of the signal by increasing or decreasing the stretching scale, and then analyze the general situation or details of the signal to realize the analysis of the local characteristics of different time scales and spaces. Wavelet threshold denoising algorithm proposed by D.L. Donoho has been widely used in data processing and other processes. At present, there are two kinds of wavelet threshold functions: hard threshold function and soft threshold function. The hard threshold function can denoise by (retaining \( \lambda \), discarding \(< \lambda \)) the wavelet coefficients, but there will be discontinuity at \( \pm \lambda \), leading to pseudo Gibbs phenomenon. On the basis of hard threshold function, soft threshold function only defines the fluctuation range of signal, which improves the continuity of the function. However, there is a certain error between the processed high-frequency signal and the original signal, which will lead to the loss of high-frequency signal and reduce the denoising effect. In order to optimize the continuity and denoising effect of the threshold function at the same time, this paper uses the soft and hard threshold compromise method to optimize the soft threshold function. By setting \( \alpha \) parameter, the threshold of wavelet is adjusted by controlling \( \alpha \) for different signals. Make up for the shortcomings of traditional functions and improve the denoising effect.

\[
\omega' = \begin{cases} 
\text{sgn}(\omega)(|\omega| - \alpha \lambda), & |\omega| \geq \lambda \\
0, & |\omega| \leq \lambda 
\end{cases}
\]

New high frequency wavelet coefficients: \( \omega' \); Decomposed high frequency wavelet coefficients: \( \omega \); Wavelet threshold: \( \lambda \); Symbolic function: \( \text{sgn} \); Regulatory factors: \( \alpha \).

3. Prediction model of air quality

3.1 data sources
In this paper, the air quality data of Chengdu, including AQI, PM2.5, PM10, SO2, Co, NO2 and O3, is selected. The sampling interval is one day, and the sampling time is from 2014 / 1 / 1 to 2020 / 7 / 29. Each group contains 2402 measurement samples, and the total number of samples is 16814. Among them, the missing data due to machine failure and other reasons are filled with the data of the next day.
Based on table 3, the impact of air pollution on the occurrence of Chengdu in different periods was analyzed and suggestions were provided.

| Air quality index | Air quality level | Pollution | Impact on travel | Measures and suggestions |
|-------------------|------------------|-----------|------------------|--------------------------|
| 0-50              | 1                | excellent | No effect        | Suitable for travel      |
| 51-100            | 2                | good      | It doesn't affect most people | A small number of people with weak resistance can reduce going out |
| 101-150           | 3                | Mild pollution | Irritable symptoms may occur in susceptible populations | Old people and children should avoid long-term outdoor sports. |
| 151-200           | 4                | Moderate pollution | Symptoms worsen and the respiratory system may be affected | People should avoid long-time outdoor sports. |
| 201-300           | 5                | Heavy pollution | People generally have dyspnea and other symptoms | The crowd should reduce going out and try to stay in the place with good indoor |
| >300              | 6                | Serious pollution | Symptoms are strongly reflected and may lead to serious diseases. | Try not to go out without an emergency. |

3.2 data preprocessing
Based on the multivariable prediction of air quality under the influence of multiple factors, the improved wavelet threshold denoising model is used to extract the air quality data. Firstly, the original data is decomposed into three levels of wavelet to get its decomposition coefficient. Then, the decomposition coefficient is thresholded based on the improved threshold and threshold function to reduce the influence of noise and realize the data denoising process. The processed coefficients are reconstructed by wavelet transform to obtain new air quality data. In this paper, three kinds of threshold functions are selected to process the noise data, and three groups of denoising sequence data are obtained.

3.3. Improved wavelet deep learning model
Due to the difference of measurement units between the data, the experimental error can be reduced effectively by normalizing it. Since the signal is decomposed by wavelet, the actual signal shows strong correlation, and the noise correlation is weak. If all wavelet coefficients are thresholded, the noise distribution in wavelet packet coefficients is not uniform, which will affect the denoising effect. In order to make up for this defect, a new improved threshold function is proposed, which calculates the correlation between the maximum decomposition scale and the wavelet packet coefficients of its adjacent scales. Taking this value as the adjustment parameter, the wavelet coefficients with the strongest positive and negative correlation are retained as far as possible, while the wavelet packet coefficients with the correlation between them are shrunk to achieve the purpose of selecting denoising. \( O_3 \) and weakly correlated PM10, SO2, Co, NO2 were denoised respectively to analyze the denoising effect.

3.4 model evaluation index
In order to evaluate the denoising effect and prediction performance of improved wavelet threshold algorithm, RNN, GRU and LSTM models for Chengdu air quality data, the mean square error and average absolute percentage error are used as the evaluation indexes of denoising and prediction performance in this experiment.

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}
\]

\[
MAPE = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{|y_i - \hat{y}_i|}{y_i} \right) \times 100\%
\]
3.5 model construction
In order to verify the effectiveness of the experimental method, the first 1802 pieces of data in each group were selected as the training data, and the last 599 pieces of data were used as the test data. The experimental environment of the program is: Windows 10 64 bit operating system, Intel (R) core (TM) i5-8300, cpu 2.30 GHz, tensorflow GPU 1.15.0, keras 2.3.1. The sequential model of keras deep learning framework is used to stack the added models. The specific process is as follows:

1. Hidden layer setting:
In order to verify the advantages and disadvantages of the model, a layer (RNN, GRU, LSTM) is added to the hidden layer of the (RNN, GRU, LSTM) model. The input data takes 1 as the time step and has seven characteristics. The number of neurons was set to 100 and 'ELU' was selected as the activation function. Add a layer of dense full junction layer and set the number of neurons to 1 to output the predicted AQI index.

2. Model parameter setting
In this paper, the batch size of the three models is set to 56, the number of iterations of the training set is 500, the average absolute error MAE is selected as the loss function, and the adaptive estimation (Adam) is used as the optimizer.

3. Analysis of experimental results
(1) After analyzing the air quality data before and after noise reduction, it can be seen that the original data fluctuates greatly, the shape is not stable enough and contains a large number of burr data. Compared with the original data, the stability of the data after wavelet threshold processing is increased, the burr data is greatly reduced, and the fluctuation of the data is weakened.(2) Compared with the original data, the improved wavelet de-noising data converges faster in depth learning training. When the iteration times are the same, the loss LSTM of the model is better than Gru and RNN in both original data and denoised data.

In order to analyze the prediction effect of the improved wavelet algorithm and different models more intuitively, the evaluation index of the model is introduced. In the wavelet algorithm, the data is decomposed into three levels, and the traditional hard threshold, soft threshold and improved soft hard threshold compromise method are selected as the threshold functions to process the decomposed data and reconstruct four groups of data. They are trained in GRU, RNN and LSTM networks respectively, and the model error is calculated. For AQI, PM2.5, PM10, SO2, Co, NO2, O3. The improved threshold denoising is carried out in 8h, and the network training is carried out. The raw air quality data were input into RNN, GRU and LSTM models respectively. We find that although the prediction effect of the model on denoised data is good, the prediction ability of the original data needs to be improved. In the process of de-noising, when the frequency bands of useful signal and noise overlap each other, the part of useful signal is removed together, which leads to the low prediction ability of the model for real data. After repeated experiments on strong correlation and weak correlation variables, we found that for AQI, PM2.5 and O3. After the denoising of Oh, the prediction effect of the model on the original data decreases sharply. However, when only PM10, SO2, Co, NO2 denoising, the actual prediction effect of the model is significantly improved. The prediction chart of LSMT model is listed as follows:

![Figure 2 model forecast and actual value](image-url)
The corresponding evaluation indexes of each model are shown in Table 2, in which a represents the use of improved wavelet threshold denoising; Y represents the use of hard threshold denoising; R represents the use of soft threshold de-noising; the following conclusions are obtained comprehensively.

1. For the prediction of air quality AQI index, PM2.5 and O3_8 concentration has a strong correlation with it. Wavelet denoising will cause the loss of useful information.

2. The original data, hard threshold and soft threshold denoised data are used for modeling. The evaluation index values of LSTM model are lower than those of GRU and RNN models.

3. The denoising algorithm based on the traditional threshold function has no obvious improvement on the prediction effect of air quality data, but the prediction effect of the deep learning model with improved wavelet threshold denoising is more significant. The prediction performance of LSTM model is about 9%, that of GRU model is 18%, and that of RNN model is 5%. It shows that the improved wavelet algorithm has better effect on air quality prediction of deep learning model.

4. The LSTM model combined with improved threshold wavelet denoising has the best performance in AQI index prediction, MAPE = 4.51%, RMSE = 4.22.

| Model | MAPE   | RMSE   |
|-------|--------|--------|
| GRU   | 6.843% | 5.567  |
| RNN   | 7.012% | 6.475  |
| LSTM  | 5.867% | 4.870  |
| A_GRU | 5.594% | 5.195  |
| A_RNN | 6.661% | 6.142  |
| A_LSTM| 5.176% | 5.314  |
| Y_GRU | 6.423% | 5.794  |
| Y_RNN | 7.017% | 6.491  |
| Y_LSTM| 6.051% | 5.950  |
| Y_GRU | 7.231% | 6.215  |
| Y_RNN | 7.054% | 6.485  |
| R_LSTM| 6.679% | 6.508  |

4. Conclusion

Through the hard threshold, soft threshold and improved soft hard threshold tradeoff method for noise reduction of Chengdu air quality data, the data of 2018 / 12 / 8-2020 / 7 / 28 were predicted by using GRU, RNN and LSTM deep learning models. The results show that due to PM2.5 and O3_8 The correlation between oh and AQI index was strong. Oh denoising will cause the loss of useful signals. The improved wavelet threshold denoising for PM10, SO2, Co, NO2 data can improve the prediction effect of the model. Compared with the traditional threshold function, the improved threshold function has better training effect on the deep learning model, and the denoised data smoothing line is better and closer to the real value. The training efficiency of the model is improved. The long-term and long-term memory neural network (LSTM) has better prediction performance in both de-noising data and original data simulation experiments.

References

[1] Zhang Xinjian, Liu Feng, Li Xiangong. Prediction of coal mine gas concentration based on wavelet denoising and cyclic neural network [J]. Coal technology, 2020,39 (09): 145-148(in Chinese)

[2] Zhang Shisen, sun Xiankun, Yin Ling, Li Shixi. GPS height time series prediction model based on wavelet denoising and neural network [J]. Global positioning system, 2019,44 (03): 117-125(in Chinese)

[3] Chen Ze, Ding Linlin, Luo Hao, song Baoyan, Zhang Ming, pan Yishan. Identification method of mine microseismic events based on improved wavelet decomposition and elm [J / OL].
Acta coal Sinica: 1-13 [2020-10-31](in Chinese)

[4] Xie Xiaoyu, Zhou junhuang, Zhang Yongjun, Wang award, Su jieying. Ultra short term renewable energy generation power prediction method based on w-bilstm [J / OL]. Power system automation: 1-14 [2020-10-31](in Chinese)

[5] Liu Bingchun, Lai Mingzhao, Qi Xin, Wang Hui. Prediction of air pollutant concentration in Beijing based on wavelet LSTM model [J]. Environmental science and technology, 2019,42 (08): 142-149(in Chinese)