CPU utilization prediction method based on composite model

Benran Hu¹, Wuzhuo Peng², Yanjun Li¹, Jiyuan Ren³, Yiqun Li¹

¹Information & Communication Branch, Heilongjiang Electric Power Company (LTD), SGCC, Harbin 150090, China;
²School of Electrical and Electronic Engineering, Harbin University of Science and Technology, Harbin 150080, China;
³Power Dispatching & Control Center, SGCC Northeast China Branch, Shenyang 110180, China
lyq300@126.com

Abstract. The business system server under the big data environment is always faced with a large number of data services, which makes the performance of its core hardware CPU face a serious test. Therefore, the prediction of server CPU utilization is of great significance in server resource allocation. In response to this situation, a method of server CPU utilization prediction based on ARMA-BiLSTM composite model is proposed. In this method, the original CPU utilization data is decomposed by wavelet decomposition, and the trend item and the detail item of CPU utilization data are obtained. Based on the Bi-LSTM neural network model, the trend item is modeled and predicted, and the detail item is modeled and predicted using the ARMA model. The prediction results of the two are added together to obtain the final prediction result. Through experimental verification, the proposed prediction model has higher accuracy than traditional models.

1. Introduction
The business system server under the big data environment generates massive data business every day, which makes the performance of the hardware resources of the business system server face a serious test. The CPU is the most overloaded part of a big data system server. Therefore, it is of great significance to predict server CPU utilization.

Since the ARIMA model was put forward, the research and application of it have never stopped. Lv et al. [1] proposed to use ARIMA model to predict European Union carbon financial market future price, so as to prevent excessive fluctuation of carbon price and stabilize the domestic market. Wei et al. [2] proposed to use ARIMA model to predict the incidence of infectious diarrhea, and achieved good results. Xu et al. [3] predicted the incidence of infection at the surgical site through ARIMA model, and the model fitting result was in good agreement with the actual value. As one of the traditional linear time series prediction methods, ARIMA is good at stationary data fitting, but it can't deal with non-stationary data fitting problems well. When the data is stationary, no difference is required, and ARMA in ARIMA is used for modeling. Since deep learning was proposed, it has received extensive attention from scholars [4].

Long short-term memory (LSTM) in deep learning has a good ability to process and predict sequences, and is widely used in time series analysis in various fields. Yao et al. [5] proposed a prediction model based on tree structure LSTM network and predicted the international gold spot
trading data in the last ten years, and achieved good results. By comparing the different topologies of dynamic neural networks and other popular parametric and non-parametric algorithms, Ma et al. [6] showed that the LSTM network can achieve the best predictive performance in terms of accuracy and stability.

Although LSTM can solve the long-term dependence problem and make up the deficiency of RNN, LSTM only uses the past information of time series data, and does not use the future information of data. Thus, the Bi-LSTM model was born [7]. Cui et al. [8] used Bi-LSTM to predict short-term traffic flow, and experiments show that it has better prediction effect. Zhang et al. [9] used Bi-LSTM to predict wheat rust, and provided scientific basis for the forecast, early warning and integrated control of wheat stripe rust.

To sum up, based on the server CPU utilization data provided by State Grid Corporation of China, this paper proposes to construct the ARMA-BiLSTM composite model by using wavelet decomposition and combining ARMA model with Bi-LSTM model, so as to predict and analyze the CPU utilization data of power grid business server. The complex CPU time series data is decomposed into trend item and detail item through wavelet decomposition, and then the trend item and the detail item are modeled by the Bi-LSTM neural network and the ARMA model. Finally, the prediction series obtained by the two are added to achieve more accurate prediction of the complex CPU utilization sequence.

2. ARMA-BiLSTM Composite Prediction Model

Because CPU utilization can be affected by virus invasion, application failure and other objective factors, with randomness and uncertainty, CPU utilization is a kind of time series data with characteristics of non-linearity, time-varying and complexity. Based on this, it is proposed to divide the original CPU utilization data into the high frequency part and the low frequency part by wavelet decomposition, and then use ARMA and Bi-LSTM to predict and superpose respectively. The specific principle is as follows.

2.1. ARMA principle

For time series with linear parts only controlled by past states, an auto regressive (AR) model can be built, with specific expression equation as shown in (1):

$$x_t = \delta + \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \ldots + \varphi_p x_{t-p} + \mu_t$$

(1)

Where \( \delta \) is a constant, and \( \varphi \) is the auto-regression coefficient. \( \mu_t \) represents white noise, which reflects the influence of external factors on the current data.

By introducing the lag operator \( L \) and making \( L X_t = X_{t-1} \), Eq. (1) can be optimized as:

$$(1 - \psi_1 L - \psi_2 L^2 - \ldots - \psi_p L^p) x_t = \delta + \mu_t$$

(2)

The characteristic equation of AR\( (p) \) is obtained, which can be expressed as:

$$\psi(L) = 1 - \psi_1 L - \psi_2 L^2 - \ldots - \psi_p L^p = 0$$

(3)

For time series whose linear part is only affected by external disturbance factors, a moving average model (MA) can be constructed, which can be expressed as:

$$x_t = \mu + \theta_1 \mu_{t-1} + \theta_2 \mu_{t-2} + \ldots + \theta_p \mu_{t-p}$$

(4)

Its characteristic equation is shown in (5):

$$\psi(L) = 1 - \theta_1 L - \theta_2 L^2 - \ldots - \theta_p L^p = 0$$

(5)

At some moments, the state quantity is affected by the past state and external disturbance at the same time. Therefore, MA term and AR term exist simultaneously in this time series model. This model is called auto-regressive moving average model, and the specific expression equation is shown in (6):

$$x_t = \psi_1 x_{t-1} + \psi_2 x_{t-2} + \ldots + \delta + \mu_t + \theta_1 \mu_{t-1} + \theta_2 \mu_{t-2} + \ldots + \theta_p \mu_{t-p}$$

(6)
2.2. Bi-LSTM principle

The Bi-LSTM model can consider both the past and future information of the data and which is expanded as shown in Fig.1. The working principle is: two hidden layer states of opposite time series are obtained through forward LSTM and backward LSTM, and then they are connected to get the same output. Forward LSTM and backward LSTM can obtain the past and future information of the input sequence respectively. At time $t$, the hidden state of Bi-LSTM $H_t$ contains forward $\tilde{h}_t$ and backward $\tilde{h}_t$:

$$
\tilde{h}_t = \text{LSTM}(h_{t-1}, x_t, e_t), t \in [1, T]
$$ (7)

$$
\tilde{h}_t = \text{LSTM}(h_{t+1}, x_t, e_t), t \in [T, 1]
$$ (8)

$$
H_t = [\tilde{h}_t, \tilde{h}_t]
$$ (9)

Where $T$ is the sequence length.

![Fig.1 Bi-LSTM network expansion diagram](image)

2.3. ARMA-BiLSTM composite model construction

First, the DB wavelet function is used to perform wavelet decomposition on the original CPU utilization data, and then the sub-sequence of CPU utilization data is reconstructed by two-interpolation reconstruction algorithm. It contains CPU utilization trend item $X_1, X_2, ..., X_N$ and CPU utilization detail item $Y_1, Y_2, ..., Y_N$.

Secondly, the trend item prediction model is constructed, and the steps are as follows:

1. CPU utilization trend item $\{X_t, t=1, 2, 3, ..., N\}$ is normalized so that the normalized data is between (-1, 1).
2. The last 60 points of the decomposed data are taken as the test data, and the rest parts are divided into the training sets.
3. Bi-LSTM uses 120 positive LSTM units and 120 negative LSTM units to form a bidirectional network model, and the number of iterations is 300.
4. The prediction result of the CPU utilization trend item data is obtained by using the data obtained in step (2).
5. Reverse normalization and restore the results predicted by the Bi-LSTM model.

Thirdly, the detail item prediction model is constructed, and the steps are as follows:

1. The autocorrelation function (ACF) and partial autocorrelation function (PACF) of the CPU utilization detail item sample data $\{Y_t, t=1, 2, 3, ..., N\}$ are calculated, and according to the obtained ACF and PACF, the model used to predict the CPU utilization detail item data is determined as ARMA ($p, q$) model.
2. The optimal model order $p$ and $q$ are determined by the obtained minimum information criterion (AIC) value.
3. Judge whether the residual sequence is a white noise sequence. If the model residual sequence is not a white noise sequence, turn step (2), modify the model again and determine the model order until the optimal model is obtained.
(4) The constructed ARMA model is used to predict the test sample set of CPU utilization detail item data in the last 60 minutes.

Finally, the trend item model prediction results and the detail item model prediction results are added together to obtain the final ARMA-BiLSTM combined prediction model result.

3. Experiment and result analysis

The experimental data used in this paper are from the sever CPU utilization data collected by the monitoring center of State Grid Corporation of China. The CPU utilization value of the business server is extracted from the monitoring software every minute, and the data of 3600 minutes is continuously extracted, and the data of 1024 minutes is intercepted as the experimental data set.

The first 964 CPU utilization data are taken as the training sample set, and the last 60 CPU utilization data are taken as the test sample set to test the prediction results.

First, different types of DB wavelet are used to decompose the original data in different layers. The decomposed data are predicted by ARMA-BiLSTM, and the error is root mean square error (RMSE). The experimental errors obtained by different DB wavelet decomposition are shown in Table 1.

| Decomposition layers | One-layer decomposition | Two-layer decomposition | Three-layer decomposition | Four-layer decomposition |
|-----------------------|------------------------|------------------------|--------------------------|------------------------|
| Wavelet type           | DB 1                   | DB 2                   | DB 3                     | DB 4                   |
|                       | 8.233                  | 7.818                  | 7.405                    | 5.592                  |
|                       | 8.551                  | 8.356                  | 7.781                    | 6.045                  |
|                       | 9.274                  | 9.611                  | 9.128                    | 6.921                  |
|                       | 10.95                  | 10.44                  | 9.840                    | 8.348                  |

From the comparison of error results in Table 1, the DB4 wavelet model has the best prediction effect. Then, according to the DB4 wavelet one-layer decomposition, two-layer decomposition, three-layer decomposition, and four-layer decomposition, RMSE is 5.592, 6.631, 7.412 and 8.515, respectively. The comparison shows that using db4 wavelet to do one-layer decomposition prediction has the best effect.

As can be seen from the error comparison results in Table 1, when DB4 wavelet is used for one-layer decomposition, the CPU utilization prediction accuracy of ARMA-BiLSTM model is the highest. Therefore, DB4 wavelet is used for one-layer decomposition of original CPU utilization, which has the best effect. The data after decomposition are shown in Fig.2 and Fig.3.

For the low-frequency trend item, the Bi-LSTM neural network modeling is used for prediction. The number of model iterations is set as 300. The prediction results are shown in Fig.4. The high-frequency detail item is a stationary time series, and AIC is used to determine the order of the model. The model is determined as ARMA \((0, 2)\), as shown in Fig.5. The final prediction result of the ARMA-BiLSTM prediction model is shown in Fig.6.
According to the feedforward neural network (BP) modeling method, the original CPU utilization data without wavelet decomposition is predicted, and the prediction result is shown in Fig.7. The same wavelet decomposition method is adopted to combine ARMA with LSTM model to obtain the ARMA-LSTM composite model, and its prediction result is shown in Fig.8.
In order to better compare the prediction results of BP, ARMA-LSTM and ARMA-BiLSTM, this paper lists the prediction errors of five prediction models, as shown in Table 2. It can be seen that the prediction accuracy of ARMA-BiLSTM is the highest, with its RMSE=5.592. The prediction accuracy of this combination model is 52.7% higher than that of BP model, and 7.49% higher than that of ARMA-LSTM composite model.

| Model        | Error |
|--------------|-------|
| BP           | 8.544 |
| ARMA-LSTM    | 6.011 |
| ARMA-BiLSTM  | 5.592 |

4. Conclusion
Aiming at the time-varying and non-linearity characteristics of power grid server CPU utilization time series, this paper proposes a composite model of server CPU utilization prediction based on ARMA-BiLSTM. The prediction RMSE of this prediction model is 5.592, and its accuracy is greatly improved compared with the traditional BP and ARMA-LSTM prediction models. This paper studies the trend development of server CPU running state and predicts the future CPU load according to its trend and threshold, so as to change the post-maintenance of power grid business server into pre-maintenance, which is of great significance to the stable operation of power grid business service system.

Acknowledgments
This work was financially supported by State Grid Heilongjiang Electric Power Company (LTD) project funding.

References
[1] Lv Jingye, Du Jingnan, et al. Forecast and enlightenment of EU carbon financial market futures price based on ARIMA model[J]. Coal Economic Research, 2019, 39(10): 23-29(in Chinese).
[2] Wei Xingang, Cao Guizhi, et al. Application of ARIMA model in the predicting incidence of other infectious diarrhea[J]. Journal of Modern Medicine & Health, 2019, 5(21): 3281-3284(in Chinese).
[3] Xu Mi, Wang Yuehui, et al. Application of ARIMA model in the prediction of incidence of surgical site infection[J]. Chinese Journal of Nosocomiology, 2020, 30(01): 141-145(in Chinese).
[4] Hinton G E. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.
[5] Yao Xiaoqiang, Hou Zhisen. Financial time series prediction by long short term memory neural
network with tree structure[J]. Journal of Computer Applications, 2018, 38(11): 3336-3341(in Chinese).

[6] Ma Xiaolei, Tao Zhimin, et al. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data[J]. Transportation Research Part C: Emerging Technologies, 2015, 54: 187-197.

[7] Alex G, Jürgen S. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5-6): 602-610.

[8] Cui Zhiyong, Ke Ruimin, et al. Stacked bidirectional and unidirectional LSTM recurrent neural network for forecasting network-wide traffic state with missing values[J]. Transportation Research Part C, 2020, 118(118).

[9] Zhang Shanwen, Wang zhen, et al. Prediction of wheat stripe rust based on knowledge graph and bidirectional long short-term memory network[J]. Transactions of the Chinese Society of Agricultural Engineering, 2020, 36(12): 172-178(in Chinese).