Electronics, Communications and Networks
A.J. Tallón-Ballesteros (Ed.)
© 2024 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA241328

A Short Term Electricity Load Forecasting Method Based on Residents in Region

Juan KONG[1], Wen-Juan SHI, Gen-Xin XIONG, Yun LI, and Qing-Ji GUO
Beijing China-Power Information Technology Co., LTD., Beijing, China

Abstract. To improve the accuracy of power load forecasting, a short-term load forecasting method based on Principal Component Analysis (PCA), XGBoost (eXtreme boosting system), and Long Short Term Memory (LSTM) using neural networks is proposed. By studying the variation patterns of residential electricity load, a feature set is constructed using date factors, climate factors, and daily load factors as inputs. Firstly, the sample dataset is divided into a load dataset and an influencing factor dataset composed of date and climate factors for data preprocessing; Then, the PCA principal component analysis method is used to extract features with significant impact, and a strong correlation feature vector is formed with the load data. The results are input into XGBoost and LSTM networks for mathematical fusion; Finally, using the electricity load data of residents in a certain area of Nanjing, the parameters of each group of network models were trained and the load prediction values were output. The results were compared and verified with the Convolutional Neural Networks (CNN) model, LSTM network model, the Gated Recurrent Neural Network (GRU) model and XGBoost LSTM network model, respectively. The results showed that the method can effectively reduce the error between the predicted values and the true values.

Keywords. short term load forecasting, principal component analysis, feature vector; XGBoost; long and short term memory

1. Introduction

In the construction of the information support system of State Grid, it is explicitly stated that research on user energy consumption data prediction technology in the field of measurement and research should be promoted [1].At present, the neural network models are commonly used in load forecasting research with multiple influencing factors, and have strong data fitting ability, making them widely used in the field of load forecasting. Reference [2] is a short-term load forecasting method for power grids based on a combination of time convolutional networks and long short-term memory networks, as well as meteorological similar day sets. It selects meteorological factors strongly related to load for load forecasting, but the prediction accuracy is not high; Reference [3] is a load forecasting method based on a hybrid neural network of convolutional neural networks and gated recurrent units, but the load forecasting feature set does not consider diverse load types; Reference [4] establishes an LSTM network structure that integrates local feature pre extraction modules and combines it in parallel with the XGBoost

[1] Corresponding author: JUAN KONG, Beijing China-Power Information Technology Co.,LTD., Beijing, China, E-mail: 15032358287@163.com.

prediction model. However, this method does not consider the LSTM network structure of feature pre extraction modules such as weather, and combines it in parallel with the XGBoost prediction model. However, this method does not consider factors such as weather that affect prediction accuracy; Reference [5] proposed a short-term load forecasting model based on XGBoost, which comprehensively considers factors such as wind speed, humidity, temperature, air pressure, and historical loads. Compared with the random forest algorithm, the accuracy has been improved, but there is still a significant gap with real data, and the amount of data is small; Reference [6] used the reciprocal error method to combine LSTM and XGBoost for ultra short term load forecasting, but there is a significant gap between the short-term load forecasting and the real dataset, and the running time is too long when considering various influencing factors.

Due to in-depth research on short-term load forecasting methods based on multiple factors, the time complexity will increase with the consideration of higher influencing factors. To ensure prediction accuracy while reducing time complexity, a neural network load forecasting method based on PCA, XGBoost, and LSTM will be proposed, taking advantage of the data dimensionality reduction and feature extraction advantages of Principal Component Analysis (PCA) [7]. The method is called PCA-XGBoost LSTM, which fully considers the multiple influencing factors of load data. The feature set constructed by date and climate factors is used as input for the PCA based feature extraction algorithm to extract important features in high-dimensional space and form high-dimensional prediction feature vectors. Combine important feature vectors with load factors and input them separately into the XGBoost based prediction algorithm and the LSTM neural network prediction algorithm for training. Mathematically fuse the prediction results, and finally output the short-term power load prediction results based on PCA-XGBoost LSTM.Using the meteorological data and electricity load data at 96 o'clock for residents in a certain area of Nanjing as samples for short-term electricity load prediction based on PCA-XGBoost-LSTM, the prediction results show that the proposed method can effectively reduce the error of load prediction and improve the accuracy of prediction.

2. Feature Set Construction and Model Introduction

2.1. Feature Set Construction

The core issue of short-term power load forecasting is to consider multiple influencing factors that affect the load and establish a load forecasting model to predict future electricity consumption behavior and frequent electricity consumption intervals of users. The insufficient sample size in the dataset affects the accuracy and robustness of the prediction model. Integrating load data with multiple influencing factors is a key factor affecting prediction accuracy. This article represents the load forecasting feature set using the following formula.

$$L(t)=L_d(t)+L_w(t)+L_n(t) \tag{1}$$

In the formula: $L(t)$ represents the actual value of the load at time t; $L_d(t)$ represents the load fluctuation caused by the date factor at time t; $L_w(t)$ represents the

load fluctuation caused by weather factors at time t; $L_n(t)$ represents the normal or trend load at time t. Based on this, this article constructs the following feature set:

(1) The date factor is an important influencing factor in short-term load forecasting. At present, the main urban electricity load in China is industrial electricity load. Numerous studies have shown that non working days (holidays, Saturdays, and Sundays) have less electricity load than working days (Monday to Friday). Build date characteristics for workdays and holidays, and $L_d(t)$ model them.

(2) Climate factors are closely related to short-term load forecasting. Among them, climate influencing factors include temperature, humidity, wind speed, weather type, etc. The power load will vary with climate changes. Therefore, meteorological influencing factor data $L_w(t)$ is constructed by collecting historical weather data, meteorological station monitoring data, and other methods.

(3) The study of load trends has strong randomness and obvious nonlinearity, and it is necessary to conduct in-depth research on the changes in power load to provide theoretical basis for load forecasting. This article utilizes the daily characteristics of load to obtain a load curve every 15 minutes, starting from 0:00 every day, to obtain the load trend $L_n(t)$.

In summary, this article takes date data, meteorological data, and historical power load data as inputs for model training, as shown in Table 1.

Table 1. Predictive feature set

Influence factor	Feature type	Feature Description
Date factor	Weekday	1 represents rest days, 2 represents working days
	Holidays	1 represents non holidays, 2 represents holidays
Meteorological factors	Maximum temperature	The highest temperature of the day
	Minimum temperature	The lowest temperature of the day
	Average temperature	The average temperature of the day
	Humidity	Daily humidity
	Rainfall	Daily rainfall
Historical load data	*L(d-n,t)*	Load value at time t in the previous n days

2.2. Principal Component Analysis

The multivariate dataset in the field of power load provides rich information for load research, but it will increase the difficulty of data collection. In many research fields, there is data correlation between variables, which adds difficulty to the research topic. If variables are analyzed separately, it is not possible to obtain the complete information contained in the data.

Therefore, there is an urgent need for a method to study the patterns of load data, while reducing the load dataset and ensuring the integrity of the original load data information, in order to achieve the goal of comprehensive analysis of the collected load data. Due to the usual correlation between load related data, closely related data is transformed into a small amount of new data, making the new data uncorrelated with

each other. A small amount of load data can represent various complete information between existing data.

If a vector $\boldsymbol{v}$ is the eigenvector of matrix $\boldsymbol{A}$, it can be represented as

$$Av = \lambda v \tag{2}$$

Among them, λ is the eigenvalue corresponding to the eigenvector $\boldsymbol{v}$.

For matrix $\boldsymbol{A}$, there is a set of eigenvectors $\boldsymbol{v}$, and by orthogonalizing this set of vectors, the orthogonal identity vector is obtained. Decompose matrix $\boldsymbol{A}$ into the following equation:

$$A = Q\sum Q^{-1} \tag{3}$$

Among them, $\boldsymbol{Q}$ is a matrix composed of eigenvectors of matrix $\boldsymbol{A}$, $\sum$ is a diagonal matrix, and the elements on the diagonal are eigenvalues.

2.3. XGBoost Model

XGBoost is an improved version of GBDT (Gradient Boosting Decision Tree). The basic component of XGboost is a decision tree, also known as a weak learner. Each decision tree is the model with the smallest objective function value, and only when the objective function value of this decision tree is the smallest, will it be selected as a weak learner. XGboost is composed of several weak learners, which can train models faster and more efficiently.

XGBoost calculates the gain before and after splitting by measuring the contribution of each leaf node to reducing overall error, in order to select the optimal splitting feature and splitting point.

$$gain = \frac{1}{2}[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}] - \gamma \tag{4}$$

Among them, $G_L = \sum_{i \in I_L} g_i$; $G_R = \sum_{i \in I_R} g_i$; $H_L = \sum_{i \in I_L} h_i$; $H_R = \sum_{i \in I_R} h_i$, I_R,I_L is the sample set of left and right subtrees. To measure the importance of features, the larger the gain value, the smaller the overall error before and after splitting.

2.4. Long Short Term Memory Network Model

LSTM belongs to a type of recurrent neural network, and its uniqueness lies in the fact that recurrent neural networks only have the function of temporary memory storage and long short-term memory. Its characteristic is the addition of input gates, forget gates, and output gates on the basis of recurrent networks. By using "gate" control, it is possible to remember long-term memory and forget unimportant information, which can effectively extract the temporal characteristics of the load. The LSTM structure diagram is shown in Figure 1, and the mathematical description is shown in Equation (5).

$$\begin{cases} f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\ i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \\ C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \\ o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ h_t = o_t * \tanh(C_t) \end{cases} \tag{5}$$

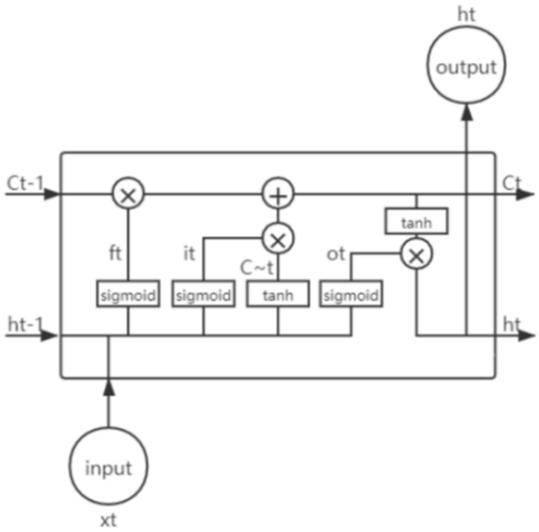

Figure 1. LSTM structure

In Figure 1 and Equation (10), f_t, h_{t-1}, h_t, x_t, b_f, b_o, b_i, b_C, i_t and o_t are respectively the forget gate, the hidden layer output at the previous time, the hidden layer output at the current time, the input vector, the bias term of the forget gate, the bias term of the output gate, the bias term of the input gate, the cell state bias term, the input gate at the current time, and the output gate at the current time. W_f, W_i, W_C, W_o respectively represent the weight matrices of forget gate, input gate, cell state, and output gate. C_t, C_{t-1}, $\tilde{C}_t$ respectively represent the cell states of the current time, previous time, and current candidate set. [] represents the connection vector; · represent matrix dot product. * represent matrix product. σ represents the sigmoid activation function. The tanh represents the tanh activation function.

3. PCA-XGBoost-LSTM Short term Load Forecasting Method

3.1. Prediction Methods

The improvement of short-term power load forecasting level depends on five aspects, namely the volume, quality, influencing factors, feature extraction, and prediction model of the predicted data. The arrival of special dates such as weather changes and holidays may lead to load fluctuations. Research has shown that historical loads, dates, meteorology, and other influencing factors can affect the accuracy of load forecasting. Therefore, when conducting power load forecasting, it is necessary to consider the impact of these factors and combine historical data for modeling and forecasting. To build a good prediction model, it is necessary to rely on a large number of historical samples for training and learning.

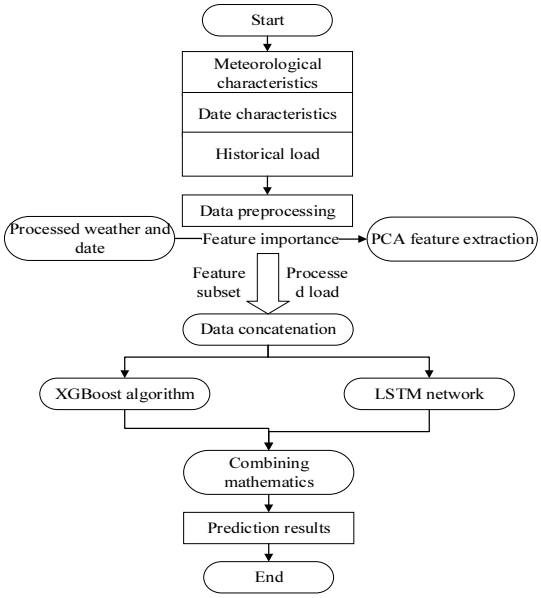

Figure 2. PCA-XGBoost-LSTM prediction method process

However, the actual collected sample data may be incomplete and inconsistent. Firstly, preprocessing operations such as data cleaning, missing value processing, and normalization will be carried out on the sample data composed of historical load, date, and meteorological data to eliminate data noise, outliers, and data mutations, thereby improving data quality. Then, by considering multiple factors that affect the accuracy of load forecasting, a PCA-XGBoost-LSTM short-term power load forecasting method is proposed by integrating feature extraction algorithms and long short-term memory network model LSTM. The process is shown in Figure 2. Input the preprocessed meteorological and date data into the PCA feature extraction algorithm to select the most influential feature data, and then concatenate it with the processed historical load data to form a new sample data. Input it into the XGBoost

algorithm and LSTM network for mathematical fusion to output the prediction results. Finally, the accuracy and efficiency of the prediction method are verified by evaluating short-term power load forecasting indicators and comparing the prediction time of various models.

3.2. Data Preprocessing

Data preprocessing process: Firstly, integrate meteorological, date, and historical load data collected from residents in the substation area; Then, the mean interpolation method is used to fill in outliers and missing values (NAN, 0, NULL, etc.) in meteorological, date, and load data, remove duplicate values, and ensure data integrity; Finally, perform normalization on the data as shown in equation (6), mapping the data to the same scale for comparison and analysis between different features. This data preprocessing technique can improve the accuracy of load forecasting models.

$$x_{new} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{6}$$

Among them, x_{new} is the normalized data; x is the sample data; $x_{\min}$ is the minimum value of the sample data; $x_{\max}$ is the maximum value of the sample data.

3.3. Specific Algorithms

The neural network prediction method of PCA-XGBoost LSTM includes three specific algorithms: PCA based feature extraction algorithm, XGBoost based prediction algorithm, and LSTM based neural network prediction algorithm. The PCA feature extraction algorithm uses meteorological and date preprocessed data as features to construct closely related new variables. Calculate the closely related feature vectors according to equations (2) and (3).

Based on XGBoost feature selection prediction algorithm, weather and date data extracted from PCA features are concatenated with preprocessed load data as input to predict power load. According to equation (4), the prediction result is calculated.

Based on the LSTM neural network prediction algorithm, meteorological and date data extracted from PCA features are concatenated with load data as input, and the LSTM algorithm is used to predict power loads. According to equation (5), the prediction result is calculated.

Finally, the prediction result G based on XGBoost feature selection prediction algorithm and the prediction result L based on LSTM neural network prediction algorithm are mathematically connected using formula (7) to obtain the power load prediction result P.

$$P = \frac{2 \times (L\varepsilon_G + G\varepsilon_L)}{\varepsilon_G + \varepsilon_L} + \frac{\varepsilon_G}{nA} \times \left(\frac{L\varepsilon_G + G\varepsilon_L}{\varepsilon_G + \varepsilon_L}\right)^2 \tag{7}$$

Among them, L is the prediction result based on the LSTM neural network prediction algorithm; G is the prediction result of the XGBoost based prediction algorithm; A is the actual power load; N represents the number of power load samples,

$\varepsilon_G = A - G$ represents the error of the prediction result G; $\varepsilon_L = A - L$ represents the error of the prediction result L.

4. Example Analysis

4.1. Experimental Data

Using the load characteristic set of residents in a certain area of Nanjing from January 1, 2012 to December 31, 2013 (time, highest temperature, lowest temperature, average temperature, humidity, rainfall, workdays, holidays, load data) as the training dataset. One day, 96 load data were collected every 15 minutes, while the other data was sampled at a 24-hour interval. The validation dataset is from January 1st to June 31st, 2014, and the testing dataset is from July 1st, 2014 to January 10th, 2015.

4.2. Evaluating Indicator

The commonly used evaluation indicators for short-term load forecasting include Mean Absolute Percentage Error (MAPE) [8], Root Mean Square Error (RMSE) [9], and Mean Absolute Error (MAE) [10]. Among them, the range of MAPE values is [0,+∞), 0% represents high-quality models, and greater than 100% represents inferior models; RMSE is an important indicator for evaluating models, and the best model can be selected; MAE is used to reflect the predictive ability of the model. The calculation formulas are shown in equations (8-10).

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - P_t}{A_t}\right| \tag{8}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(A_t - P_t)^2} \tag{9}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|A_t - P_t| \tag{10}$$

Among them, At is the true value of the sample; Pt is the predicted value of the sample; n is the number of samples.

4.3. Result Analysis

Use the PCA-XGBoost-LSTM short-term load forecasting model proposed in this article, along with CNN model, LSTM model, GRU model, and XGBoost-LSTM combination model, to fit the sample training set data. Preprocess the test set according to formula (5) in section 2.2 to improve the quality of the predicted data, and then select the corresponding model for power load forecasting. Using the above prediction model, calculate the average absolute percentage error (MAPE), root mean square error (RMSE), and average absolute error (MAE) on three types of test samples. The comparison results are shown in Table 2.

Table 2. Comparison of prediction model accuracy

Prediction Model	MAPE/%	RMSE/MW	MAE/MW
CNN	5.32	434.55	387.46
LSTM	2.91	277.93	212.43
GRU	4.03	287.54	251.72
XGBoost-LSTM	4.06	360.20	296.43
PCA-XGBoost-LSTM	1.63	141.64	110.30

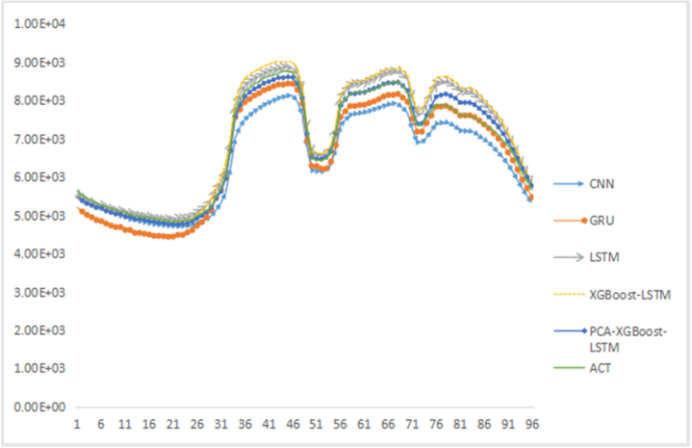

Figure 3. Comparison between predicted results of various models and actual values

The experimental results showed that MAPE decreased by 3.69%, 1.28%, 2.4%, and 2.43%, respectively; RMSE decreased by 67%, 49%, 51%, and 61% respectively; MAE decreased by 72%, 48%, 56%, and 63% respectively, confirming that the PCA-XGBoost LSTM prediction method has high accuracy in load forecasting.

In order to visually observe the accuracy of each model, this article predicts the power load curve of each model based on the data at 96 o'clock on January 10, 2015 in the test set, and compares it with the actual power load curve (ACT). The comparison results between the predicted results of each model and the actual values are shown in Figure 3. The PCA-XGBoost-LSTM short-term load forecasting model proposed in this article has a high fitting degree with the actual load curve and has good prediction accuracy.

5. Conclusion

A short-term load forecasting method based on regional residents at the district level was proposed. Firstly, data preprocessing methods were used to clean, process missing values, and normalize the collected historical meteorological, date, and load data. Then, the feature extraction algorithm based on PCA selected the feature vectors with greater influence, combined with the preprocessed load data, and used as inputs for the XGBoost

based prediction algorithm and the LSTM neural network prediction algorithm. Finally, the prediction results of the two algorithms were mathematically fused to output the load prediction value.

The established load forecasting feature set does not fully consider the influencing factors, such as electricity prices, economic factors, social factors, etc. Further research will be conducted on the impact of multiple influencing factors on load forecasting, in order to further improve the accuracy of power load forecasting.

References

[1] Kang Chongqing,Xia Qing,Zhang Boming. Review of power system load forecasting and its development[J].Automation of Electric Power Systems,2004,28(17):1-11(in Chinese).

[2] Liu Hui,Ling Ningqing,Luo Zhiqiang,et al.A short-term load forecasting method for power grids based on TCN-LSTM and meteorological similar day sets [J]. Shaanxi Electric Power,2022 (008):050.

[3] Yao Chengwen et al. Load forecasting method based on CNN-GRU hybrid neural network [J]. Power Grid Technology,2020,44 (9).

[4] Zhuang Jiayi,Yang Guohua,Zheng Haofeng,et al.A CNN LSTM XGBoost short-term power load forecasting method based on multi model fusion [J].China Electric Power,2021,54 (5):46-55.

[5] Shen Yu,Xiang Kangli,Huang Xianan,et al.Research on short-term load forecasting based on XGBoost algorithm [J].Water Resources and Hydropower Technology,2019 (S1):6.

[6] Chen Zhenyu,Liu Jinbo,Li Chen,et al.Ultra short term power load forecasting based on the combination model of LSTM and XGBoost [J].Grid Technology,2020 (2):7.

[7] Huang Xiangjun.Application of BP neural network based on principal component analysis in power system load forecasting [J].Technology Information, 2008 (16):2.

[8] Dong Jiafu,Wan Xiong,Wang Yan, et al. Short term power load forecasting based on XGB Transformer model [J]. Power Information and Communication Technology, 2023,21 (1): 10.

[9] Ouyang Fulian,Wang Jun,Zhou Hangxia. Short term power load forecasting method based on improved transfer learning and multi-scale CNN Bilstm Attention [J]. Power System Protection and Control, 2023, 51 (2): 9.

[10] Zhao Qian,Zheng Guilin. Short term power load forecasting based on WD-LSSVM-LSTM model [J]. Electrical Measurement and Instrumentation, 2023, 60 (1): 6.