Attention Based Spatial-Temporal Graph Convolutional Networks for Short-term Load Forecasting

Rong Liu 1,*, Luan Chen 1

1 School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu 611731, Sichuan Province, China

*Corresponding author’s e-mail: 201921040224@std.uestc.edu.cn

Abstract. To predict the load of the power system with a known network structure, this paper proposes a novel attention based spatial-temporal graph convolutional network (ASTGCN) model to predict the node load in the power grid. The experimental results show the good performance of ASTGCN.

1. Introduction

Power system load forecasting means that the power load is based on its own changes as well as from the influence of economic and meteorological factors. Load forecasting is roughly divided into short-term load forecasting, mid-term load forecasting and long-term load forecasting in terms of time. Among them, short-term load forecasting mainly predicts the load from one day to one week [1].

Short-term load forecasting has great significance: first, it provides guarantee for the safe and economic operation of the power system; second, it provides a basis for grid scheduling, power supply, and transaction plans in a market environment; third, it provides a basis for users to manage power consumption plans [1]. The effect of short-term load forecasting depends on the accuracy of the forecast. Recently, studies have begun to publish load data together with the grid structure. This information makes it impossible to ignore the spatial information when predicting the load. Therefore, this article mainly studies short-term load forecasting with grid structure, and discusses the accuracy of node load forecasting.

2. Related work

In the past, the methods of load forecasting mainly included time series method, expert system method, wavelet analysis method, fuzzy prediction method [1], etc. However, because the short-term load is affected by a variety of random factors [2], such as the randomness of users and the complexity of the weather, the application range and prediction accuracy of many methods are limited.

Recently, people have begun to use neural networks to predict load. Literature [3] developed an improved deep residual network to improve the short-term coincidence prediction results. In addition, more scholars began to combine multiple models for load forecasting. Literature [4] combines convolutional neural network (CNN) and long short-term memory (LSTM), CNN-LSTM neural network can extract complex load features to effectively predict the load.

With the development of graph convolutional network models, graph convolutional neural networks can be used to capture the structural features of graph networks. Literature [5] proposed a time graph convolutional neural network (T-GCN), which integrates graph convolutional networks and gated recurrent units for tasks based on urban road traffic prediction.
The attention mechanism is widely used because it can highlight important data \cite{6}, especially in time series forecasting. To predict crowd flow, literature \cite{7} uses the regional-level attention mechanism to learn the time shift of each region in the long-term periodic dependence. The prediction results show that the model can improve the crowd when the time shift varies with the region.

This paper proposes a spatial-temporal graph network model based on the attention mechanism to predict load with grid structure. Temporal and spatial attention mechanism to dynamically capture the time and space correlation of load data. The spatial-temporal graph convolution module is used to describe the time and space characteristics of the load data.

3. Attention based Spatial-Temporal Graph Convolutional Networks

The main framework of the ASTGCN model is as follows. The input mainly has three temporal components, which are hour component, day component and weekly component. After these three temporal components is the same space-time block. The spatial-temporal block is composed of spatial-temporal attention mechanism and spatial-temporal convolutional network. Next is the fully connected layer separately. Finally, we use the weights of the three components obtained by training to combine the predicted values of the three components to obtain the final predicted value.

3.1. The Input series

Firstly, we need to consider the impact of past consecutive hours on the future load. Secondly, as the periodicity of the day, the same moment of each day has a certain similarity. Finally, we need to consider the periodicity of the week, there are similarities in the same time with the week as the cycle. These constitute our three temporal components, namely the hour component, the day component and the week component.

3.2. Temporal and spatial attention mechanism

The model contains two attention mechanisms, temporal attention mechanism and spatial attention mechanism. The spatial attention mechanism is used to adaptively capture the dynamic correlation between nodes in the spatial dimension \cite{8}. Taking the spatial attention mechanism of the hour part as an example:

\[
S = V_s \sigma \left( X_h^{(r-1)} W_1 W_2 (X_h^{(r-1)})^T + b_s \right)
\]  

(1)
where $\chi_h^{(r-1)} = (X_1, X_2, ..., X_{r-1})$ is the input of the spatial-temporal block. $V_S, b_x, W_1, W_2, W_3$ are the parameters obtained through training, and $\sigma$ represents the activation function. The element $S_{i,j}$ in the matrix $S$ represents the strength of the association between node $i$ and node $j$. The softmax function is used to ensure that the sum of the attention weights of the nodes is $1$, and the matrix $S'$ is normalized to obtain $S'$.

Similarly, use the attention mechanism to consider the importance of different times:

$$E = V_e \sigma\left(\left(\chi_h^{(r-1)} \right)^T U_1 \left(U_2 \chi_h^{(r-1)} + b_e\right)\right)$$

(3)

$$E'_{i,j} = \frac{\exp(E_{i,j})}{\sum_{j=1}^{N} \exp(E_{i,j})}$$

(4)

among them, $V_e, b_e, U_1, U_2, U_3$ are the parameters obtained through training. The value of element $E_{i,j}$ in $E$ represents the correlation between time $i$ and $j$. Finally, the softmax function is also used to normalize the matrix $E$, and the normalized time attention matrix $E'$ is obtained [8].

3.3. Spatial-temporal convolution

The spatial-temporal convolution module includes graph convolution in the spatial dimension and standard convolution in the time dimension.

(1) graph convolution in spatial dimension

The Laplacian matrix of the graph is defined as $L = D - A$, where $A$ is the adjacency matrix, and $D \in \mathbb{R}^{N \times N}$ is the diagonal matrix, which is composed of node association degrees $D_{ii} = \sum_j A_{ij}$. The eigenvalue decomposition of the Laplace matrix is $L = U \Lambda U^T$, where $\Lambda = \text{diag}\{[\lambda_1, ..., \lambda_N]\} \in \mathbb{R}^{N \times N}$ [8].

Using the convolution theorem, we can multiply the signal in the spectral space, and then use the inverse fourier transform to convert the signal to the original space to achieve graph convolution. Acting the convolution kernel $g_\theta$ on the signal $x$, the graph convolution can be expressed as the following form

$$x^* G y = U \left((U^T x) \odot (U^T y)\right) = U g_\theta (\Lambda) U^T x = g_\theta^* G x$$

(5)

Among them, $^* G$ represents the graph convolution operator, $x, y$ represent the signal of the node domain on the graph, $\odot$ refers to the Hadamard calculation method, and $^* x = U^T x$ represents the Fourier transform of the signal [8].

However, when the scale of the graph is large, it is difficult to directly perform eigenvalue decomposition on the Laplacian matrix. To solve this problem, this paper adopts Chebyshev polynomial [8].

$$g_\theta^* G x = g_\theta (L) x = \sum_{k=0}^{K-1} \theta_k T_k (L) x$$

(6)
among them, the parameter $\theta \in R^k$ is a vector of polynomial coefficients. $\bar{L} = \frac{2}{\lambda_{\text{max}}} L - I_N$, $\lambda_{\text{max}}$ is the maximum eigenvalue. The cycle of Chebyshev polynomials is defined as $T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$, where $T_0(x) = 1, T_1(x) = x$. The graph convolution model uses ReLU as the final activation function [8]. To dynamically adjust the correlation between nodes, for each term of the Chebyshev polynomial, $T_i(\bar{L})$ is combined with the spatial attention matrix $S' \in R^{N \times N}$, the above graph convolution calculation becomes [8]:

$$g_\theta^* G x = g_\theta^* (L) x = \sum_{k=0}^{K-1} \theta_k \left( T_k(\bar{L}) \odot S' \right) x$$

(7)

(2) Convolution in temporal dimension
After the graph convolution operation, a standard convolutional layer in the time dimension is also superimposed to update the information of the node by merging the information on the adjacent time slices [8]. Also take the $r^{th}$ layer in the hour component as an example:

$$x_h^{(r)} = \text{ReLU} \left( \Phi \ast \left( \text{ReLU} \left( g_\theta^* G x_h^{(r-1)} \right) \right) \right) \in R^{C \times N \times T}$$

(8)

* represents the standard convolution operation, $\Phi$ is the time-dimensional convolution kernel parameter, and ReLU is the activation function.

3.4. Multi-component fusion
The hourly, daily, and weekly components of different nodes have different effects on the load forecast output of the node. We need to learn their respective weights from historical data. Therefore, the final prediction result after fusion is

$$\hat{Y} = \hat{Y}_h \odot \hat{Y}_d + \hat{Y}_d \odot \hat{Y}_w + \hat{Y}_w \odot \hat{Y}_w$$

(9)

$W_h$, $W_d$ and $W_w$ are learning parameters, $\hat{Y}_h, \hat{Y}_d, \hat{Y}_w$ are the output results of the three time blocks respectively.

4. Experiment

4.1. Experiment data
The experimental data set is real data from the United States. The load data of 240 nodes in 2017 was measured with an electric meter, collected every hour, and it also includes 134 connection relationships of these 240 nodes [9]. We delete the nodes whose load is 0 all the time. Divide all data into a ratio of 6:2:2 as training set, validation set and test set.

4.2. Experiment metrics
Use the metrics MAE, RMSE, and MAPE to evaluate the accuracy of the network’s prediction, and calculate the variance of MAPE, namely SMAPE, to measure the stability of the algorithm. For all of them, lower value is better. We first calculate the metrics of each node, and then average the metrics to get the metrics of the entire system.

4.3. Baselines and parameter settings
The comparison models we have chosen are BPNN, LSTM, GRU, MSTGCN. Compared with ASTGCN, MSTGCN just has no attention mechanism. The parameters of the model are optimizer=‘adam’, learning rate=0.001, epochs=50, batchsize=32. For these parameters, all models are the same. Especially, for the ASTGCN model, predicting the load in the next hour, the input sequence is the sequence one week ago, one day ago, and one hour ago, and the time point of the week and day of the sequence is the same as the predicted time point.
4.4. Experimental results and analysis
The following table shows the index values of all model load forecasting. The index values of ASTGCN are the smallest, followed by MSTGCN.

Table 1. The predicted results of the model

| model  | MAE(KWh) | RMSE(KWh) | MAPE(%) | SMAPE  |
|--------|----------|-----------|---------|--------|
| BPNN   | 1.69     | 2.26      | 32.23   | 0.1584 |
| LSTM   | 1.65     | 2.23      | 28.02   | 0.1233 |
| GRU    | 1.62     | 2.19      | 27.45   | 0.1243 |
| MSTGCN | 1.31     | 1.89      | 21.70   | 0.0805 |
| ASTGCN | 1.27     | 1.81      | 21.39   | 0.0789 |

Visually display the values of the three indicators MAE, RMSE, and MAPE of the models. On the whole they are consistent with the average of the three indicators.

The experimental results show that the ASTGCN prediction node load is not only accurate but also stable. Compared with MSTGCN, it reflects the advantages of time attention mechanism and spatial attention. In addition, the difference between ASTGCN and MSTGCN is far less than that between them and BPNN, LSTM and GRU. It shows that when forecasting the load with graph structure, the spatial-temporal convolution module which can capture the spatial and temporal characteristics is more advantageous than the model which can only capture the time feature.

5. Conclusion
This paper proposes a spatial-temporal graph convolutional neural network model based on the attention mechanism, and applies the model to load forecasting. The experimental results show that the performance of the ASTGCN model is better than other models. The model is used to predict the load
at the regional level, which is different from only predicting the node load in time dimension. The disadvantage is that when the grid topology is too large and the connection relationship is too complex, the graph neural network needs to be further optimized to adapt to more complex networks.

Acknowledgments
This work was supported by Sichuan Science and Technology Program (NO. 2018HH0146).

References
[1] Khan I, Hafeez G, Alimgeer K S, “Electric Load Forecasting based on Deep Learning and Optimized by Heuristic Algorithm in Smart Grid”, Applied Energy, 269(2020).
[2] Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. 2018. Modeling Long- and Short-Term Temporal Patterns with Deep Neural Networks. In Proceedings of ACM Conference (SIGIR’18). ACM, New York, NY, USA, 11 pages.
[3] K. Chen, K. Chen, Q. Wang, Z. He, J. Hu and J. He, “Short-Term Load Forecasting With Deep Residual Networks,” IEEE Transactions on Smart Grid,10(4),3943-3952, (2019).
[4] Kim T Y , Cho S B, “Predicting Residential Energy Consumption using CNN-LSTM Neural Networks.” Energy,182(SEP.1),72-81(2019).
[5] L. Zhao et al., “T-GCN: A Temporal Graph Convolutional Network for Traffic Prediction,” IEEE Transactions on Intelligent Transportation Systems, 21, (9), 3848-3858, (2020).
[6] Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. “Show, attend and tell: Neural image caption generation with visual attention”, International conference on machine learning, 2048–2057, (2015).
[7] Tian, Chujie, Zhu, Xinning, Hu, Zheng, and Ma, Jian. “Deep Spatial-temporal Networks for Crowd Flows Prediction by Dilated Convolutions and Region-shifting Attention Mechanism.” Applied Intelligence, (Dordrecht, Netherlands) ,50.10,3057-070, (2020).
[8] Guo, Shengnan, Lin, Youfang, Feng, Ning, Song, Chao, and Wan, Huaiyu. “Attention Based Spatial-Temporal Graph Convolutional Networks for Traffic Flow Forecasting.”, Proceedings of the AAAI Conference on Artificial Intelligence 33: 922-29, (2019).
[9] F. Bu, Y. Yuan, Z. Wang, K. Dehghanpour, and A. Kimber, “A Time-series Distribution Test System based on Real Utility Data.”, 2019 North American Power Symposium (NAPS), Wichita, KS, USA, pp. 1-6(2019).