Short-term Load Forecasting Model Considering Meteorological Factors

Xi Yunhua¹, Zhu Haojun¹, Dong Nan¹

¹China Southern Power Grid Energy Development Research Institute Co., Ltd., 510000 Guangzhou, China

Abstract: Because of the limitation of basic data and processing methods, the traditional load characteristic analysis method can not achieve user-level refined prediction. This paper builds a user-level short-term load forecasting model based on algorithms such as decision trees and neural networks in big data technology. Firstly, based on the grey relational analysis method, the influence of meteorological factors on load characteristics is quantitatively analyzed. The key factors are selected as input vectors of decision tree algorithm. This paper builds a category label for each daily load curve after clustering the user's historical load data. The decision tree algorithm is used to establish classification rules and classify the days to be predicted. Finally, Elman neural network is used to predict the short-term load of a user, and the validity of the model is verified.

0 Introduction

Load forecasting is the basis of power system planning, operation, demand side management and other work. Due to the limitations of basic data and processing methods, traditional load forecasting methods rarely achieve user-level refined prediction. Based on the large data technology, it is important to establish the relationship between meteorological factors and load forecasting, excavate the inherent law of load characteristics, and construct an accurate short-term load forecasting model for the economic operation and planning of power grid.

Typical load forecasting methods include time series^[1-2], support vector machine^[3], neural network^[4], and combination forecasting^[5]. Literature^[6] Short-term load forecasting is realized by establishing autoregressive moving average model (ARMA model), and the stability of power load is analyzed by non-parametric test method. Literature [7] reduces the amount of data in the sample by looking for historical short-term loads similar to the meteorological characteristics of the day to be predicted. The fuzzy classification is used to fuzzyly assign meteorological factors that have a great influence on the electrical load, so as to determine the historical load data required for the prediction. The SVM is used to predict the grid load in Shanxi in the short term. Compared with the BP neural network, the accuracy of the prediction is significantly improved. Document [8] combines chaotic characteristics of load and least squares support vector machine to select the nearest neighbouring points of the prediction point as the training sample data of least squares support vector machine, the prediction accuracy is relatively high. Some researchers [9] - [12] shorten the

training time and improve the accuracy of the prediction model by improving the neural network prediction model.

In this paper, a refined short-term load forecasting model considering meteorological factors is established by means of clustering analysis, grey correlation analysis, CART decision tree and Elman neural network algorithm. Based on grey relational analysis, key meteorological factors are selected as input vectors of decision tree algorithm. Cluster analysis is used to establish classification labels for load curves, and classification rules are established through CART decision tree. Finally, Elman neural network is used for short-term load forecasting to realize fine load forecasting for a user.

1 Selection of influencing factors of load characteristics

Grey relational degree analysis is a method to evaluate the selected objects after considering various factors comprehensively. The relational relationship is determined by comparing the similarity between reference sequence and comparison sequence. The more similar the geometric shape of reference sequence is to that of comparison sequence, the greater the correlation degree obtained by grey correlation analysis. Based on the results of grey correlation analysis, the main factors affecting load characteristics can be selected.

1.1 Basic steps of grey relational degree analysis

1.1.1 Determining comparison sequence and reference sequenc

The first step is to determine the reference sequence (systematic feature sequence) and comparison sequence (influence factor sequence) that need to be used in the quantitative analysis process. Let $Y = (y_1 \ y_2 \ \cdots \ y_t)^T$ be the reference sequence, y_i ($i = 1, 2, \dots, t$) be the value of the reference sequence, and t the total amount of data. The influencing factors $X_1, X_2, \dots, X_n$ which form the comparison sequence matrix can be expressed in the following forms:

$$(X_1, X_2, \dots, X_n) = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1t} & x_{2t} & \cdots & x_{nt} \end{bmatrix}_{t \times n} \quad (1)$$

In the above formula, n denotes the number of impact factors that need to be evaluated in the whole evaluation process.

1.1.2 Normalizing Sequence Matrix

The numerical values of different sequences are quite different, and the dimensions of different types of data are not comparable, which will affect the accuracy of the final analysis results, so it is necessary to standardize the sequence. In this section, the maximum value is used to normalize the data set. The calculation formula is as follows:

$$\begin{cases} \bar{x}_{ij} = \frac{x_{ij}}{\max_{i=1,2,\dots,t} (x_{ij})} \\ \bar{y}_i = \frac{y_i}{\max_{i=1,2,\dots,t} (y_i)} \end{cases} \quad i = 1, 2, \dots, t \quad j = 1, 2, \dots, n \quad (2)$$

The normalized reference sequence (system feature sequence) and the comparative sequence (influencing factor sequence) are respectively:

$$\bar{Y} = (\bar{y}_1 \ \bar{y}_2 \ \cdots \ \bar{y}_t)^T \quad (3)$$

$$(\bar{X}_1, \bar{X}_2, \dots, \bar{X}_n) = \begin{bmatrix} \bar{x}_{11} & \bar{x}_{12} & \cdots & \bar{x}_{1n} \\ \bar{x}_{21} & \bar{x}_{22} & \cdots & \bar{x}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{x}_{t1} & \bar{x}_{t2} & \cdots & \bar{x}_{tn} \end{bmatrix}_{t \times n} \quad (4)$$

1.1.3 Generating difference matrix

Considering reference sequence and comparison sequence as points in multidimensional space, we can study the relationship between various factors and system characteristics in a specific t -dimensional space. The calculation formula of difference sequence is as follows:

$$\begin{aligned} \Delta_j &= |X_j - Y| = (|x_{j1} - y_1|, \dots, |x_{jt} - y_t|)^T \\ &\stackrel{\text{define}}{=} (\Delta_{j1}, \dots, \Delta_{jt})^T \quad j = 1, 2, \dots, n \end{aligned} \quad (5)$$

Then the difference matrix Δ is:

$$\Delta = \begin{bmatrix} \Delta_{11} & \Delta_{12} & \cdots & \Delta_{1n} \\ \Delta_{21} & \Delta_{22} & \cdots & \Delta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \Delta_{t1} & \Delta_{t2} & \cdots & \Delta_{tn} \end{bmatrix}_{t \times n} \quad (6)$$

1.1.4 Computing the correlation matrix between the comparison sequence and the reference sequence

The correlation coefficient reflects the correlation between the characteristic sequence of the system and the sequence of influencing factors at different time points. The calculation formula is as follows:

$$\ell_{ji} = \frac{\min_{j=1,\dots,n} \min_{i=1,\dots,t} \Delta_{ji} + \rho \max_{j=1,\dots,n} \max_{i=1,\dots,t} \Delta_{ji}}{\Delta_{ji} + \rho \max_{j=1,\dots,n} \max_{i=1,\dots,t} \Delta_{ji}} \quad (7)$$

$$i = 1, \dots, t \quad j = 1, \dots, n$$

In the above formula, $\rho \in (0, 1)$ is called resolution coefficient. The smaller the resolution, the stronger the resolution. Then the correlation coefficient matrix is obtained as follows:

$$\ell = \begin{bmatrix} \ell_{11} & \ell_{12} & \cdots & \ell_{1n} \\ \ell_{21} & \ell_{22} & \cdots & \ell_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \ell_{t1} & \ell_{t2} & \cdots & \ell_{tn} \end{bmatrix}_{t \times n} \quad (8)$$

1.1.5 The results of grey correlation analysis are obtained.

In order to calculate the correlation degree between the system characteristic sequence and the subsequence of each influencing factor, the correlation coefficients at each time point are weighted and summated, as shown in the following formula:

$$L_j = \sum_{i=1}^t \omega(i) \ell_{ji} \quad (9)$$

In the formula above, L_j is the correlation degree between the feature sequence of the system and the j -th influencing factor and $\omega(i)$ is the weight of the correlation coefficient at the i -th time point are considered.

1.2 Analysis of the effect of meteorological factors on load characteristics

Taking a certain user as the research object, the daily load characteristics of the user's historical power consumption are analyzed, including daily maximum load, daily minimum load, daily peak-valley difference, daily peak-valley difference, daily minimum load rate, and the correlation between daily minimum load rate and meteorological factors such as temperature (maximum, minimum, average), daily average humidity, air pressure (maximum, minimum, average), daily maximum wind speed, etc. The key factors affecting the user's electricity consumption characteristics are obtained to prepare the

input vectors of the decision tree in the next section. The results of grey correlation analysis are shown in the following table:

Table 1 Grey Relevance Analysis of User Load Characteristics

	T1	T2	T3	T4	T5	T6	T7	T8
P1	0.75	0.56	0.54	0.65	0.56	0.73	0.65	0.55
P2	0.74	0.59	0.61	0.74	0.59	0.73	0.68	0.59
P3	0.72	0.54	0.56	0.68	0.54	0.70	0.63	0.54
P4	0.80	0.66	0.49	0.58	0.66	0.78	0.80	0.66
P5	0.80	0.60	0.50	0.60	0.59	0.78	0.70	0.59
P6	0.76	0.64	0.50	0.61	0.64	0.74	0.76	0.64
P7	0.65	0.53	0.65	0.80	0.53	0.64	0.62	0.53

In this table, T1, T2, T3, T4, T5, T6, T7, T8 represents daily maximum temperature, daily maximum barometric pressure, Maximum daily wind speed, daily minimum temperature, daily minimum barometric pressure, daily average temperature, daily average humidity, daily mean barometric pressure respectively. And P1, P2, P3, P4, P5, P6, P7 represents daily peak load, daily minimum load, daily peak-valley difference, daily peak-valley difference rate, daily average load, daily load rate and minimum daily load rate respectively.

As can be seen from the above table, temperature has the greatest and most direct impact on user load characteristics. Daily maximum temperature, daily minimum temperature, daily average temperature and daily average humidity are the four meteorological factors that have more than 0.6 grey correlation degree with most of the characteristics of daily load, which have a greater impact on the user's electricity load. Therefore, this paper chooses the daily maximum temperature, daily minimum temperature, daily average temperature and daily average humidity as the historical electricity load based on decision tree. Input vector for classification.

2 classification of predicted daily load based on CART decision tree

Decision Tree (DT) is an analysis method which constructs a decision tree to obtain the probability that the expected NPV is greater than or equal to zero on the basis of known sample data occurrence probability. According to the different types of input data and prediction objectives, it can be divided into classification tree and regression tree: classification tree can achieve the classification prediction of unknown types of objects; regression tree can achieve the prediction of continuous values.

In this paper, CART decision tree is used to classify the user's historical electricity data and to establish classification rules. The idea is as follows:

Fig. 1 The process of establishing user load classification rules

In the figure above, the middle part is the CART decision tree algorithm model, and the left part is the influencing factor set of the user's daily load curve input by the decision tree algorithm. As the sample eigenvalue of the daily load, the right part is the user's power consumption category obtained by clustering analysis using k-means algorithm. As the class label of the sample, the sample eigenvalue and the sample label are input into the decision tree. By training in the model, the classification rules of electricity load can be established, and the sample eigenvalues (meteorological factors) of the days to be predicted can be input into the established classification rules, and the classification results of the days to be predicted can be obtained. Finally, the same type of daily data sets as the days to be predicted can be found.

3 Load forecasting model based on elman neural network and CART decision tree

Elman neural network was proposed by J.L. Elman in 1990, which is a dynamic regression neural network. Elman neural network added a receptive layer as a delay operator in the hidden layer, which has the function of short-term memory. Therefore, it has the ability to process dynamic information and can recognize and detect time-varying patterns.

3.1 Elman neural network structure

Elman neural network includes input layer, hidden layer, acceptance layer and output layer. The structure of the whole network is as follows:

Fig. 2 Elman neural network structure

Elman neural network uses the sum of squares of errors as learning index function to determine model parameters.

3.2 Load forecasting based on elman neural network and CART decision tree

Elman neural network has a stronger dynamic memory function than the traditional static feedforward neural network, so it is suitable for building load forecasting model of time series. The flow chart of Elman neural network prediction is as follows:

Fig. 3 Elman neural network prediction

4 Case study

In the analysis of this example, firstly, we use clustering algorithm to classify the user's historical electricity data and get the category label of the historical data; secondly, grey relational degree analysis algorithm is used to get the meteorological factors which have great influence on the user as the input vector of the decision tree algorithm; secondly, input the selected meteorological factors and category labels into the decision tree to get the classification rules and to forecast. The meteorological factors of the day are input into the classification rules established by the decision tree, and the classification results of the day to be predicted are obtained. Finally, the historical electricity data of the same kind of day to be predicted are used as the training set, and the daily load curve of December 31, 2016 is predicted by Elman neural network.

4.1 Cluster analysis of historical electricity data

In this section, the K-means clustering algorithm is used to classify the daily load curves of users from January 2015 to December 2016. Each daily load curve is composed of 24 full-time active power loads per day. The clustering results of the users are as follows:

Fig. 5 Clustering results of user history load

4.2 Selection of key influencing factors for users

In this paper, the correlation between meteorological factors such as temperature, humidity, air pressure, maximum wind speed and load characteristics has been quantitatively analyzed by using grey correlation analysis. The meteorological factors which have great influence on daily load characteristics have been selected, and the results are shown in the following table:

Table 2 Selection of Key Influencing Factors of Load Characteristics

	T1	T4	T6	T7
P1	0.7487	0.6522	0.7257	0.6504
P2	0.7404	0.7426	0.7253	0.6789
P3	0.7204	0.6797	0.7005	0.6297
P4	0.7968	0.5791	0.7816	0.8009
P5	0.8004	0.5998	0.7758	0.6984
P6	0.7626	0.6061	0.7412	0.7628
P7	0.6532	0.8011	0.6387	0.6187

The four meteorological indices of daily maximum temperature, daily minimum temperature, daily average temperature and daily average humidity are selected as the main meteorological influencing factors and become the input vectors of the follow-up decision tree algorithm.

4.3 Classification of predicted days based on decision tree

Based on the preparation of the previous two steps, the user's historical meteorological data, i.e. daily maximum temperature, daily minimum temperature, daily average temperature and daily average humidity, are input into the CART decision tree algorithm as sample eigenvalues, and the class labels of the sample data obtained by clustering analysis are also input. The classification rules shown in the following figure are established:

Fig. 6 Classification Rules Established by Decision Tree

In the figure above, x1 represents the daily maximum temperature, x2 represents the daily minimum temperature, x3 represents the daily average temperature, x4 represents the daily average humidity, a total of 15 nodes. The values of meteorological factors on the day to be predicted are shown in the following table:

Table 3 Values of meteorological factors for predicted days

T1	T4	T6	T7
16.3°C	9.4°C	13.2°C	69.6%

Based on the classification rules established by decision tree algorithm, the meteorological factors of the day to be predicted are input, and the day to be predicted is divided into category 1.

4.4 Load forecasting based on elman neural network

Table 4 Load curve forecasting results table based on this forecasting method

time	1:00	2:00	3:00	4:00	5:00	6:00	7:00	8:00	9:00	10:00	11:00	12:00
Real value (MW)	0.245	0.237	0.294	0.251	0.201	0.185	0.176	0.184	0.273	0.334	0.372	0.369
predicted value(MW)	0.260	0.241	0.277	0.234	0.199	0.194	0.182	0.176	0.263	0.315	0.349	0.375
error(%)	6.16	1.79	5.78	6.73	0.94	4.92	3.46	4.24	3.59	5.80	6.32	1.66
time	13:00	14:00	15:00	16:00	17:00	18:00	19:00	20:00	21:00	22:00	23:00	24:00
Real value (MW)	0.338	0.329	0.318	0.259	0.336	0.421	0.431	0.463	0.428	0.333	0.298	0.264
predicted value(MW)	0.328	0.322	0.319	0.241	0.311	0.413	0.436	0.440	0.437	0.341	0.281	0.255
error(%)	3.08	2.07	0.41	6.81	7.64	1.78	1.03	4.95	2.18	2.49	5.70	3.27

Based on the preparation of the first three steps, Elman neural network can be used for load forecasting. The third step results show that the day to be predicted belongs to the first category of clustering results. Therefore, the historical load of the category label belonging to the first category is used as the training set of Elman neural network. The final load prediction results are shown in the following figure:

Fig. 5 Load prediction results

In the figure above, curve ① represents the actual value, curve ② represents the direct forecasting value, curve ③ represents the Elman neural network method forecasting value.

The average relative error of the daily load curve obtained by the method described in this paper is 3.87%. The load data of two years before the forecast date are trained directly by Elman neural network. The average relative error of the forecasting result is 12.53%. The method in this paper reduces the error of the load curve forecasting obviously and improves the accuracy of the load curve forecasting. Although the accuracy of the user-level load forecasting is improved, the load curve is not smooth due to the large fluctuation of the daily load curve of the user level. There are a few errors in the load forecast for 24 hourly points of the day, resulting in a final error of 3.87%. Overall, this method has reference value for user-level load forecasting.

5 Summary

This paper obtains the category label of user's historical daily load curve based on clustering analysis; obtains four key factors influencing user's daily load characteristics based on grey correlation analysis: daily maximum temperature, daily minimum temperature, daily average temperature and daily average humidity; selects the same type of daily data set based on CART decision tree and the day to be predicted; and finally carries out load pre-loading based on Elman neural network. Measurement. A refined short-term load forecasting model considering meteorological factors is established by synthesizing various methods in this paper.

Compared with training historical load data sets directly using Elman neural network to obtain prediction results, the method used in this paper chooses the same type of daily data sets as the day to be predicted, which improves the prediction accuracy. The refined load forecasting model established in this paper has reference and research value for user-level load forecasting, and can also be extended to load forecasting of regional power grids (provincial and municipal power grids).

References

1. Zhang Kaikai, Yanping, Elec Swit. **54** 91-96 (2016)
2. Fan Jinji, ZJ Elec Power, **8** (2018)
3. Yang Licheng, GXU (2008)
4. Zhang Lan. J NWU, **46** 639-644 (2016)
5. Jin Liang. JLU (2018).
6. Tang Liangyan. SCUT (2010).
7. Niu Dongxiao, Gu Zhihong, Xing Mian. Wang Huiqing. CSEE,, **26** 6-12 (2006)
8. Yang Zhuo. XAUT (2008).
9. Xu Xing. NCU (2012).
10. Yang Long. CSU (2009).
11. Sun Xincheng, Kong Jianshou, Liu Zhao. J NJUST, **42** (2018)
12. Meng Anbo, Hu Hanwu, Liu Xiangdong. Power Sys Prot Ctrl, **44** (2016)