Short-term Load Prediction Based on the Combination of K-means and Random Forest

Weiwei Chai¹, Lei Qin², Haiying Dong¹,²*, Chunshan Sun¹, Wei Shen¹ and Shiyun Qiao¹

¹ Lanzhou New Area Power Supply Company, State Grid Gansu Electric Power Comp- any, Lanzhou, Gansu, China
² School of Automation & Electrical Engineering, Lanzhou Jiaotong University, Lanzhou, Gansu, China
Email: hydong@mail.lzjtu.cn

Abstract. Aiming at the problem that the power supply and distribution system runs at low load rate for a long time and wastes capacity due to the expansion of the power supply and distribution system, a short-term load forecasting method combining K-means and random forest is proposed. The proposed method divides power users into four categories based on electricity behavior, based on which the corresponding category load data is selected as the input sample of the random forest model to obtain short-term load prediction results. Example analysis shows that this method can ensure the rapid clustering accuracy, and effectively realize the short-term prediction of power load based on the random forest, to achieve the purpose of improving the load rate.

Keywords: K-means algorithm; Low load rate; Load forecasting; Power supply and distribution scheme; Random forest;

1. Introduction
Reasonable power supply and distribution scheme is conducive to the economic operation of power system and ensures the operating efficiency of power company. In the process of optimizing the regional electricity sales mode, Lanzhou New Area Power Supply Company found that the power grid was in a state of low load rate after users’ access. In view of this problem, it is important to guide the power supply and distribution scheme using the load forecast in line with the local actual conditions.

The effect of the load clustering directly affects the accuracy of the prediction, and a good load clustering is the prerequisite of the load prediction. Literature [1] proposes a fusion load curve clustering algorithm based on the wavelet transform for the problem that the existing clustering algorithm is not suitable for large-volume, high-dimensional load curve clustering. Literature [2] proposes a new load curve clustering method using image processing technology, which simplifies the understanding of power consumption patterns and makes some improvements. Literature [3] proposes methods based on deep learning and improved rapid partition of grid reactive voltage based on the K-means algorithm, improving the grid voltage regulation problem. Based on aviation user information, literature [4] builds a user recognition technology based on K-means algorithm and BP neural network, and accurately clusters and predicts row users.

At present, short-term power load prediction methods are mainly divided into traditional prediction methods and artificial intelligence prediction methods [5]. Traditional prediction methods are mature and simple and still being used. Intelligent learning algorithm based on neural network and other
methods has self-learning and adaptive function; literature [6] proposes a new statistical load prediction to obtain the actual graph of the resulting risk of the load demand profile in view of the lack of expected risk information about the uncertainty of the predictor in the traditional method. Literature [7] proposes a rolling load prediction method combining the clustering algorithm with the random forest, which significantly improves the generalization ability of the load prediction and is robust. Literature [8] realizes classification, regression and time series models based on random forest, which solves the problem that current load prediction is difficult to simulate the distribution of electricity load. The random forest algorithm has strong generalization ability, easy overfitting, and insensitive noise sensitivity significantly improve the load prediction accuracy.

Based on the above study, this paper proposes a short-term load prediction method combining clustering and prediction to solve the problem of low load rate. This method reduces the clustering time with K-means algorithm, Based on random forest, short-term power load prediction is used to guide power system scheduling, so as to improve the load rate.

2. The K-means Clustering Principle
The K-means algorithm measures the similarity of data by calculating the distance between data objects. The algorithm belongs to an unsupervised clustering algorithm, and the main principle is to divide the sample set with N data samples into k classes according to the rules, and the internal data samples are highly similar [9]. Less similarity between the different classes.

Euclidean distance between the data samples [10]. The formula is shown in Equation (1).

\[ d = \sum_{i=1}^{k} \sum_{u \in C_i} \left\| u - w_i \right\|_2^2 \]  

Where, \( w_i \) is the central sample of cluster \( C_i \).

The mean of the k clusters was calculated as the new cluster center, as shown in Equation (2).

\[ w_i = \frac{1}{C_i} \sum_{u \in C_i} u \]  

The termination condition is set to \( |w_{n+1} - w_n| \leq \epsilon \).

Euclidean distance was used to describe the similarity of the data samples [11], after classification, correlation coefficient (3) is used to indicate the similarity between data [12]-[13].

\[ P_{xy} = \frac{\sum_{i}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i}(x_i - \bar{x})^2 \sum_{i}(y_i - \bar{y})^2}} \]  

Where, \( x_i \) and \( y_i \) are the ith attribute value of \( x \) and \( y \) respectively.

3. Load Prediction Based on K-means Clustering with Random Forest

3.1. Principles of Random Forest Regression
The core idea of random forest regression is to generate the corresponding CART decision tree from the training set of group K randomly using the Bootstrap resampling method [14], the results for each decision tree were averaged for random forest predictions.

The minimum mean variance was used as the basis for the optimal node splitting [15]:

\[ H(x_m) = \frac{1}{N_m} \sum_{i=N_m}^{y} (y - \bar{y}_m)^2 \]  

Where, \( \bar{y}_m \) and \( X_m \) are the average value and training sample set of sample target features of m node.

Due to limited data. This paper uses the correlation between load data as a feature set and includes two features: the predicted moment load and the load close to the previous moment point; the forecast moment load at the same time as the past few days \([7]\). Thus the feature set is shown in formula (5).

\[
X_i = \{(x_{0,i-n}, x_{0,i-n+1}, \ldots, x_{0,i-1}, x_{m,j-1}, x_{m-1,j}, \ldots, x_{1,1})\}
\]

\( x_i, x_{m,n} \) in formula, the feature set of the predicted i ti- me; the load data at the n moment of day m from the predict- ed time.

3.2. Short-term Power Load Prediction Model Based on Random Forest

K-means algorithm is used to classify the annual load data and then random forest is used to predict the short-term load, as shown in figure 1.

3.2.1. Raw annual load data were clustered by K-means.
3.2.2. The category to be tested was selected according to the clustering results, and the corresponding load data sample was selected as the random forest input sample.
3.2.3. Data samples were divided into training set D and test set S, where the training set D sample size was N and the eigenvector M dimension.
3.2.4. A k sub-training set of sample size N was collected using Bootstrap resampling. The k decision tree is generated according to the CART algorithm, and each decision tree selects the m-dimensional features from the M-dimensional feature vectors that randomly fail to put back to the split nodes, traverses the feature vectors and selects the optimal split criterion based on the mean square difference.
3.2.5. The test set S is input into the random forest, and the final prediction results are averaged for each decision tree according to the weight data generated by the test set, as weighted by Equation (6).

\[
y = \sum_{i=1}^{K} w_i Y_i
\]

4. Example Analysis

4.1. Data Base

30 different user load data of Lanzhou New Area in 2020 were taken as cluster samples. Select the
category to which the sample belongs to be tested, input the random forest model from January 2020 to September 2020 as the training set, and make load prediction in October 2020 as the test set.

Figure 2 shows the load curve of 30 users in 2020:

![Load curve for 2020.](image)

The average absolute percentage error is used as the evaluation index, as shown in formula (7):

\[
\text{MAPE} = \frac{1}{N} \sum_{t=1}^{K} \left| \frac{y_t - x_t}{x_t} \right| \times 100\%
\]

Where, \( x_t \) is the actual value, \( y_t \) is the predicted value.

4.2. Example Analysis
First, data processing is carried out to delete duplicate data and vacant values [13]. The processed data were read in as the initial sample data. Analysis figure 3 of the number of cluster and cluster relationship curve shows that the distance between the data samples significantly reduced selection with the number of clusters \( n = 4 \).

![Number of clusters and cluster relation curves.](image)

The 2020 power loads were clustered using the K-means algorithm, and the clustering results are shown in figure 4, it can be seen that the second class load and the fourth class load are significantly higher than the other two loads. From January to February 2020, the cliff decline affected by the epidemic in 2020, we can see that this load is industrial load.

The first type of load reaches the maximum annual load from September. The addition of heating load in winter causes this fluctuation and meets the characteristics of residential electricity load.

The third class of load did not fluctuate significantly throughout the year, the load value is low and the load curve is smooth in line with the local agricultural load characteristics.
Datasets were selected from industrial load as load prediction samples, with a total of 263 days of daily sampling frequency of 1 h from January-September 2020 data as training set and October data as test set for random forest load prediction, and prediction results are shown in figure 5.

Short-term load forecasting is used to forecast the power load of one day to one week in the future. It is mainly used to arrange daily or weekly scheduling plans to ensure the safe, economic and stable operation of the system.

The error index values based on different predicted days are shown in table 1. The elevated error values with the predicted days were significantly based on local actual prediction at over 6 days based on a significant error increase of the method.

| Predicted days | MAPE   |
|---------------|--------|
| 5 days        | 7.96   |
| 6 days        | 7.78   |
| 7 days        | 14.21  |
| 8 days        | 15.23  |
| 9 days        | 19.29  |
| 10 days       | 21.67  |
| 15 Days       | 21.96  |
| 30 Days       | 29.03  |

The experimental verification for the small problems can be based on random forest short-term load forecast time of no more than 6 days can help the power department reasonably arrange the generator set start and stop, choose the operation mode of power system, realize the power system economic scheduling, so as to ensure the safety, stability and economic operation of power system, improve the load rate of power system.
5. Conclusion
In this paper, we propose to combine K-means clustering with random forest to predict the power load in Lanzhou New Area. The example analysis shows the following conclusions:

(1) Lanzhou New Area has distinct load characteristics and obvious differences between different loads. The advantages of simple and fast K-means clustering can make accurate clustering of regional load.

(2) Combining K-means clustering with random forest, using post-clustering data as predictive samples can accurately predict industry-specific short-term load trends. The prediction results show that the method has high confidence at the prediction time exceeding 6 days and has certain reference value for optimizing the power supply and distribution scheme.

Acknowledgments
This project is funded by the State Grid Lanzhou new area power supply company of Gansu Province.

References
[1] Jiang Z, Lin R, Yang F and B Wu 2018 A fused load curve clustering algorithm based on wavelet transform IEEE Transactions on Industrial Informatics 14: 1856-1865.
[2] Park K and Son S 2019 A novel load image profile-based electricity load clustering methodology IEEE Access 7: 59048-59058.
[3] Zhao J J, Jia R, Ling H, Zhu T T 2021 Research on fast partition of reactive voltage in power grid based on deep learning and improved k-means clustering algorithm Power System Protection and Control 49: 89-95.
[4] Zhu F, Wang Y Q 2021 Research on user information clustering and prediction based on K-means and neural network machine learning algorithm Intelligence Science 39: 83-90.
[5] Li Y, Jia Y J, Li L, Hao J Sh, Zhang X Y 2020 Short-term power load prediction based on the Random Forest algorithm Power System Protection and Control 48: 117-124.
[6] Aprillia H, Yang H T and Huang C M 2021 Statistical load forecasting using optimal quantile regression random forest and risk assessment index IEEE Transactions on Smart Grid 12: 1467-1480.
[7] Xun G Y 2018 Based on cluster analysis with random forest for short-term load rolling prediction Smart City 4: 9-11.
[8] Li W H, Chen H, Guo K, Guo S R, Han J M, Chen Y Zh 2016 Research on power load prediction based on stochastic forest algorithm Computer Engineering and Applications 52: 236-243.
[9] Huang Zh Q 2020 Multi-Label Classification and Label Completion Algorithm Based on K-Means Anqing Normal University.
[10] Li B Ch, Yu H J, Liu J Y 2021 Short-term PV generation power prediction based on K-means and CEEMD-PE-LSTM Hydropower and Energy Science 39: 204-208.
[11] Xue L Y 2018 K-means Algorithm in Geological Disaster System Xi'an University of Technology, China.
[12] Li Ch H 2020 Study on Short-term Power Load Forecasting Based on Machine Learning Algorithm University of Electronic Technology, China.
[13] Chen L D, Xu Ch, Zhao X, Pan Y, Zhang Y J 2020 Analysis of typical industry and user load characteristics in Guizhou Southern Energy Construction 7: 29-35.
[14] Huang H, Sun K, Liu D 2018 Study on hourly load forecclition of power system based on random forest Smart Power 46: 8-14.
[15] Xing Sh H 2020 Research on the Short-Term Load Prediction of an Electric Power System Based on the Random Forest Algorithm Qingdao University, China.