Prediction and Analysis of Short-Term Load Forecasting Model Based on Similar Day Clustering and CatBoost

Zhonghua Chen1*, Jiamin Chen1, Caiqian Wang1

1 Hangzhou Electric Power Design Institute Co., Ltd., Hangzhou, Zhejiang, ZIP code, Country China

*Corresponding author’s e-mail: zhonghua1217@163.com

Abstract. Accurate load forecasting provides a reference of vital importance for power generation and power dispatching systems. A short-term load forecasting model combining similar day clustering and CatBoost method is proposed in this paper. Similar day clustering is a method for evaluating the similarity between the days to be forecast and historical days. Similar day clustering is based on K-means clustering which can classify samples with similar features into the same category. Then, CatBoost is used to build load forecasting models. CatBoost enables us to directly handle categorical features without pre-processing. This paper utilized the information collected by a substation in Hangzhou as the data set, and compare the proposed model with the existing forecasting model based on Autoregressive Moving Average, Gradient Boosting Decision Tree, CatBoost, Long Short-Term Memory. The experimental results show that the proposed model provides more accurate results compared with other models in terms of MAPE, RMSE.

1. Introduction
The stability of the power industry is very important to the development of society. In order to ensure the reliability and efficiency in operation of the power system and to meet the power needs of users, it is necessary to make reasonable plans for power generation and power dispatch [1]. Accurate load forecasting will provide an important reference for power generation and power dispatch. In this paper, we are mainly discussing short-term load forecasting in units of days.

Short-term load is affected by many factors, including weather, date, and month. The influence of these factors on the load is very complicated, and we need to take a large number of factors into consideration when building a forecasting model [2].

ARMA, as a traditional short-term load forecasting method, was proposed in 1970. ARMA is a time series analysis method that predicts the future load based on the trend changes of historical load data. ARMA is a simple and easy-to-use prediction method, but it is not good at predicting non-stationary time series, and it lacks analysis of the impact of various factors on load. Therefore, this method has greater limitations, which must be taken into consideration.

Support vector machine (SVM) method has been successfully applied to various fields as an important machine learning algorithm. One of the characteristics of SVM is using kernel functions to convert nonlinear cases into linear cases. SVM is suitable for processing high-dimensional data and has good generalization ability [3].

Che proposed a load forecasting model based on SVM algorithm which utilized a combination of kernels, and achieved higher prediction accuracy than the single-kernel SVM model [4]. When building SVM models, additional attention should be paid to data preprocessing and parameter tuning, which will have a great impact on the efficiency and accuracy of the model.
Gradient Boosting Decision Tree (GBDT) algorithm builds the model by combining lots of weak base learners, that is, decision trees [5]. GBDT algorithm runs multiple iteration, for each iteration, each decision tree is trained based on the residuals of the current model. GBDT is suitable for cases with multiple influencing factors and does not require too much parameter tuning efforts to get good forecasting accuracy. CatBoost is a novel learning algorithm based on GBDT, open sourced by Yandex in 2017 [6]. The substantive character of CatBoost model enables it to integrate some innovative and effective algorithms that can handle classification features and reduce over-fitting problems, thereby improving the accuracy of the model [7].

When building a short-term load forecasting model, in addition to considering the impact of various factors on the load value, we can also consider obtaining references directly from historical loads, mining the similarity between the features of the forecasting day and the features of the historical days. It is predictable to see that, generally speaking, two samples with similar features have similar label values. Similar day clustering is one such method by classifying the samples with similar features into the same category to find the connection between the forecasting days and the historical days [8]. Similar day clustering can be understood as a pre-training process, which can not only provide a reference for subsequent model predictions, but also speed up the convergence speed of the model training [9].

This paper proposes a short-term load forecasting model combining similar day clustering and CatBoost method. The algorithm principles of similar day clustering and CatBoost are discussed in Section II. Section III reviews the structure setup and details of the model. In Section IV, the experiment results of the proposed model are shown and compared with models based on GBDT, LSTM, CatBoost.

2. Methodology

2.1. CatBoost

2.1.1. Capability to Deal with Categorical Features. Traditional GBDT algorithm can’t deal with categorical features, which means that if you want to use categorical features in GBDT models, you need to process the categorical features in advance. The most widely used pre-processing method is one-hot encoding, but the features may become sparse after one-hot encoding. However, CatBoost can handle categorical features by transform categorical features to numerical features, which is called ordered target statistic. Ordered target statistics can estimate the expected target value of each category in an effective way [6]. First, CatBoost randomly arranges the dataset. Then CatBoost calculates the average label value for each sample using the samples that are located before the sample and has the same categorical value as the sample. CatBoost computes the ordered target statistics according to the average label value using the following formula:

$$x_k^\prime = \frac{\sum x_{j\in D_k} [x_j^\prime = x_k^\prime] \cdot y_j + \beta P}{\sum x_{j\in D_k} [x_j^\prime = x_k^\prime] + \beta}$$

In this formula, \( D = \{(x_k, y_k)\}_{k=1,...,n} \) means the dataset, \( x_k = (x_k^1,...,x_k^n) \) means the feature vector of the \( k \)-th sample, \( y_k \) means the corresponding label value, \( [x_j^\prime = x_k^\prime] \) equals 1 if \( x_j^\prime \) equals \( x_k^\prime \), otherwise equals 0. For the \( k \)-th sample of the training dataset, \( D_k \) means the dataset before the given sample in the random permutation. For the \( k \)-th sample of the test dataset, \( D_k = D \), which means the whole training dataset. The role of \( P \) and \( \beta \) is to help smooth the noise for low-frequency categories.

2.1.2. Feature Combinations. CatBoost uses combinations of categorical features as additional features that capture high-order dependencies. When the tree seeks for a new split, CatBoost consider in a greedy manner where CatBoost combines all combinations and categorical features [7].
with GBDT, CatBoost can achieve higher prediction accuracy and lower overfitting.

2.2. Similar Day Clustering
Similar day clustering method is based on K-means clustering. The purpose of similar day clustering is to add a day-category feature to all samples and to classify the samples with similar features into the same category. Each category is numbered and the number is used as the value of the day-category feature.

The steps of similar day clustering are: first, the features of all sample are uniformly normalized, and then K-means clustering is used to classify the samples with similar distances into the same category, where Euclidean distance of the features is used to evaluate the distance between the samples [10].

K-means clustering is essentially an iterative solution clustering analysis algorithm. K-means divides the dataset into k groups, each group has a cluster center, and each sample is assigned to the cluster center closest to it [11]. Use K-means clustering to classify the samples, and the final goal is to add a category feature called day-category feature.

Similar day clustering algorithm can evaluate the similarity between forecasting day and historical days, and use samples with similar features to provide references for the forecasting days. Similar day clustering is equivalent to a pre-training method and can improve the accuracy of subsequent forecasting models.

3. Model Overview
This proposed short-term load forecasting model based on similar day clustering and CatBoost is illustrated in Fig. 1.

First, extract some preliminary features. The features that can be extracted include: (1) Time features, i.e. the year, the quarter, the month, the week, the day, the day of the year, the day of the week, weekday, weekend; (2) Weather features, i.e. various statistical values of temperature and humidity of the day, the statistical values include the maximum, minimum, average and standard deviation [12]; (3) Holiday features, indicating whether this day is in holiday.

Then, use similar day clustering to add day-category feature to all samples to form a complete dataset. Classify the samples so that samples with similar features are classified into the same category.

Finally, train CatBoost model and use CatBoost model to make predictions. Categorical features can be directly used without one-hot encoding [13].

4. Experiment Results
This paper utilized the load data of a substation in Hangzhou covering one year as the label set, and use day as the sample unit. The data between December 11th and December 20th of the year is used as the label.
set of training set, and the remaining data is used as the label set of test set. In this experiment, we extracted daily weather features. The weather data is recorded in each hour, mainly recording temperature and humidity values. Based on these weather data, we calculated three temperature and humidity statistics for each sample, including the maximum, minimum and average values.

The Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE) are utilized as reference models to validate the accuracy of the proposed method. MAPE focuses on the relative error of the test set and RMSE focuses on the standard deviation of the absolute error of the test set [14].

The MAPE is defined as:

$$MAPE = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{y_i - \bar{y}_i}{\bar{y}_i} \right| \times 100\%$$

(2)

where $N$ means the number of samples of the data set, $y_i$ means the forecasting value of the $i$-th sample, and $\bar{y}_i$ means the label value of the $i$-th sample.

The RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{y}_i)^2}$$

(3)

where the meanings of the symbols are the same as the meanings of the symbol in equation 2.

We compared the forecasting accuracy of the proposed model with existing models in terms of MAPE, RMSE. The forecasting scores of all the model are shown in Table 1.

Table 1. Comparison of the Performance of Proposed Model with Existing Models

| Model               | MAPA      | RMSE    |
|---------------------|-----------|---------|
| ARMA                | 6.88%     | 3.211   |
| GBDT                | 4.73%     | 2.0469  |
| CatBoost            | 2.82%     | 1.2758  |
| LSTM                | 3.47%     | 1.4678  |
| Similar day+CatBoost| 2.10%     | 0.9069  |

According to the results of Table 1, it can be seen that the accuracy of the proposed model is higher than ARMA, GBDT, CatBoost, LSTM in terms of MAPE, RMSE. Furthermore, the forecasting result of these models are plotted in Fig. 2. And the outcome of the proposed model fits the actual load best among these models.

Figure 2. Load forecasting results
5. Conclusion
In this paper, we propose a hybrid model for short-term load forecasting, which is based on similar day clustering and CatBoost. First, we extract various types of features including categorical features and numeric features. Secondly, normalize the features and use similar day clustering to classify samples with similar features into the same category. Number each category and use the number as the value of the newly added day-category feature, which can be used to measure the similarity between forecasting day and historical days. Afterwards, the CatBoost model is trained as the load forecasting model. CatBoost can directly handle categorical features without preprocessing, which is useful for dealing with load forecasting cases since load forecasting cases usually have many categorical features. Comparing to ARMA, GBDT, CatBoost and LSTM model, the proposed method achieves a higher forecasting accuracy than other methods, using the same data set in terms of MAPE, RMSE.

References
[1] Shyh-Jier-Huang and Kuang-Rong Shih, “Short-term load forecasting via ARMA model identification including non-Gaussian process considerations,” IEEE Transactions on Power Systems, 18(2): 673-679, 2003.
[2] Kim, K.-H.; Park, J.-K.; Hwang, K.-J.; Kim, S.-H. Implementation of hybrid short-term load forecasting system using artificial neural networks and fuzzy expert systems. IEEE Trans. Power Syst. 1995, 10, 1534–1539.
[3] Belousov A, Verzakov SA, Von Frese J. A flexible classification approach with optimal generalisation performance; support vector machines. ChemomIntell Lab Syst 2002;64:15-25
[4] Che J X, Wang J Z. Short-term load forecasting using a kernel-based support vector regression combination model. Applied Energy, 2014, 132(1): 602-609.
[5] L.Y. Xiao, W. Shao and M.X Yu, “Research and application of a hybrid wavelet neural network model with the improved cuckoo search algorithm for electrical power system forecasting,” Applied Energy, 198: 203-222, 2017.
[6] A.V. Dorogush, V. Ershov, and A. Gulin, “CatBoost: gradient boosting with categorical features support,” arXiv preprint arXiv:1810.11363, 2018
[7] L. Prokhorenkova, G. Gusev, A.Vorobev, A.V.Dorogush and A.Gulin “CatBoost: unbiased boosting with categorical features,” Advances in Neural Information Processing Systems, pages 6637-6647, 2018.
[8] S. Li, P. Wang, and L. Goel, “A novel wavelet-based ensemble method for short-term load forecasting with hybrid neural networks and feature selection,” IEEE Transactions on Power Systems, vol. 31, no. 3, pp. 1788–1798, Jun. 2016.
[9] H. Hahn, S. Meyer-Nieberg, and S. Pickl, “Electric load forecasting methods: Tools for decision making,” European Journal of Operational Research, vol. 199, no. 3, pp. 902–907, Dec. 2009.
[10] M. Qi, K. Guolin, W. Taifeng, C. Wei, Y. Qiwei, M. Weidong, L. TieYan, "A Communication-Efficient Parallel Algorithm for Decision Tree," Advances in Neural Information Processing Systems, vol. 29, pp. 1279-1287, 2016.
[11] A. R. Reis and A. A. Da Silva, “Feature extraction via multiresolution analysis for short-term load forecasting,” IEEE Trans. Power Syst., vol. 20, no. 1, pp. 189–198, Jan. 2005.
[12] A. K. Jain, “Data clustering: 50 years beyond K-means,” Pattern Recognit. Lett., vol. 31, pp. 651–666, Jun. 2010.
[13] S. Fan, L. Chen, and W.-J. Lee, “Short-term load forecasting using comprehensive combination based on multimeeteorological information,” IEEE Trans. Ind. Appl., vol. 45, no. 4, pp. 1460–1466, Jul./Aug. 2009.
[14] Y. Goude, R. Nedellec, and N. Kong, “Local short and middle term electricity load forecasting with semi-parametric additive models,” IEEE Trans. Smart Grid, vol. 5, no. 1, pp. 440–446, Jan. 2014.