Electric Power Systems Research 190 (2021) 106841

Contents lists available at ScienceDirect

Electric Power Systems Research

journal homepage: www.elsevier.com/locate/epsr

A hybrid prediction model based on pattern sequence-based matching method and extreme gradient boosting for holiday load forecasting

Kedong Zhu*, Jian Geng, Ke Wang

Power Automation Department, China Electric Power Research Institute, Nanjing 210003, China

ARTICLE INFO

Keywords:
Holiday effects
Pattern sequence-based matching method
Public holidays
Short-term load forecast
Time-series decomposition
XGBoost

ABSTRACT

In short-term load forecast (STLF), forecasting holiday load is one of the most challenging problems. Aimed at this problem, a hybrid prediction model based on pattern sequence-based matching method and extreme gradient boosting (XGBoost) is presented. It divides holiday STLF problem into the predictions for proportional curve and daily extremum of electricity demand, which are relatively independent and relate to different factors. It is benefit for holiday STLF by task decomposing. Based on the shape similarity measured by Euclidean distance, the proportional curve is predicted by pattern sequence-based matching method. Daily extremum of electricity demand is predicted by XGBoost considering holiday classification. Finally, the predicted holiday load profile is synthesized from the above two prediction results with segment correction. The proposed methodology can analyze holiday load characteristics more effectively and get a higher prediction accuracy independent of sufficient data and expert experience. We evaluate our methodology with many algorithms on a real data set of one provincial capital city in eastern China. The results of case studies show that the proposed methodology gives much better forecasting accuracy with an average error 2.98% in holidays.

1. Introduction

Short-term load forecast (STLF) plays a decisive role in maintaining the security and economy of power system. As a complex nonlinear problem, STLF is closely related to the social changes, economic factors and weather change [1,2]. Various methods have been applied to improve the forecasting accuracy, such as linear auto-regressive (AR) model, neural network (NN), fuzzy technology, support vector machine (SVM), etc. [3-7]. With the relatively sufficient data, the methods mentioned can get satisfactory prediction results on the normal days. However, holiday load is usually atypical and can create significant predicted error easily, which is viewed as a challenging problem in STLF.

Recently, an enormous amount of research effort goes into improving holiday STLF techniques. Ref. [8,9] present a state space forecast method with multiple seasonal patterns and a forecast method with double seasonal auto-regressive moving average mechanism, respectively. Because of the restriction of linear statistical models, they are only built based on load data without other factor data. In Ref. [10,11], a Mahalanobis distance based fuzzy polynomial regression method and a fuzzy expert system with similar day method are presented for holiday STLF problem. Ref. [12,13] focus on the expert rules that are summarized from historical data and used in recognizing the similar mode. Although good results are achieved, the principal difficulty is the generation for many fuzzy/expert rules required. In Ref. [14,15], the similar day methods are applied, which usually select the historical samples due to the similarity with some typical factors to sort. However, more different factors will have different impacts on holiday load so that the selected historical similar sample is hard to reflect the comprehensive effect of all holiday factors. Ref. [16,17] adopt NN model in holiday STLF problem and get the proper results. But NN performance are strongly related to model parameters, which are often determined by experience. More importantly, the number of holiday samples to train NN is usually insufficient.

Until now, it is still challenging to predict the holiday load accurately, mainly in the following reasons: (1) Compared with normal day, holiday load is more complicated and has no adequate and effective data. Thus, conventional intelligent algorithm trained by large samples is difficult to achieve a high precision. (2) Different holidays vary in the load shape and load base. It is because all the factor effects on holiday load are quite different. Therefore, it is so hard for conventional analysis to excavate the hidden laws that the characteristic analysis of holiday load profile is often given by expert experience.

In view of the problems above, this paper presents a novel hybrid prediction model based on pattern sequence-based matching method and extreme gradient boosting (XGBoost). It divides holiday STLF

[*] Corresponding author.
E-mail addresses: hhuzkd@sina.com (K. Zhu), gengjian@epri.sgcc.com.cn (J. Geng), wangke@epri.sgcc.com.cn (K. Wang).

https://doi.org/10.1016/j.epsr.2020.106841
Received 21 April 2020; Received in revised form 13 July 2020; Accepted 1 September 2020
Available online 15 September 2020

problem into the predictions for proportional curve and daily extremum of electricity demand. Through the shape similarity measured by Euclidean distance, the former is obtained from pattern sequence-based matching method. After holiday classification, the latter is obtained from XGBoost trained by samples with regard to holiday effect and other common effects. Then, the predicted holiday load profile is synthesized from the above two prediction results and is corrected in the segment of early morning. The main contribution of this paper is to convert the hidden law mining into the obvious law mining by task decomposing in holiday STLF problem. The proposed methodology can evaluate and estimate different factor effects on holiday load independent of sufficient data and expert experience. Through experiment comparison, the proposed methodology can be applied to holiday STLF problem due to its wide applicability. Its application in improving the quality and accuracy of holiday STLF is novel.

In the next section, we discuss the holiday load characteristics. Section 3 presents the hybrid prediction model based on pattern sequence-based matching method and XGBoost. Section 4 provides the experimental results. Section 5 concludes this paper.

2. Holiday load characteristics

Holidays generally belong to the specific days when the behavioral changes happen due to social institutions and national policy [18,19]. For discussing holiday load characteristics, the hourly electricity demand of one provincial capital city in eastern China is used in this section. Respectively from two essential properties, proportional curve and daily extremum of electricity demand are further analyzed [20]. Proportional curve describes the load shape, standing for load operation mode while daily extremum of electricity demand represents the load base.

2.1. The characteristic analysis for load profile

Fig. 1 shows the hourly electricity demand in 2015. The days plotted by green circle represent New Year Day, Spring Day, Qingming Festival, May Day, Dragon Boat Festival, Mid-Autumn Festival and National Day, respectively. It can be found that some downward trends happen during holidays.

Fig. 2 presents the load profiles of different holidays and normal days. It can be found that holiday load profiles are different from the normal days in load shape and load base. Compared with normal weekday, the power-intensive industrial load is reduced much in holiday, which leads to a huge difference in load profile. Although the composition of holiday load (e.g. residence and service industry) is similar to that of normal weekend, there still exists some differences in load profile owing to holiday tourist impact by national policy.

2.2. The characteristic analysis for proportional curve

Fig. 3 plots the load profiles of New Year day across three years. Load base has been increasing year after year but their load shapes are similar on the whole. For further analysis, proportional curve is extracted from the linear mapping of load profile based on daily extremums. The formula of proportional curve is as follows:

$$p_{h,t} = \frac{P_{h,t} - \min(P_{h,t})}{\max(P_{h,t}) - \min(P_{h,t})} \tag{1}$$

where $P_{h,t}$ and $p_{h,t}$ represent the hourly electricity demand and hourly load percentage corresponding to a certain holiday h, respectively.

Fig. 4 shows the proportional curves of New Year day during the same period. These proportional curves are very similar, only varying slightly in the medium load period and evening peak period. It means that there exists the similarity in holiday load composition.

2.3. The characteristic analysis for daily extremum of electricity demand

According to holiday length, two kinds of holidays usually happen in China, 3-days holiday and 7-days holiday. New Year Day, Qingming Festival, May Day, Dragon Boat Festival and Mid-Autumn Festival belong to 3-days holiday while Spring Day and National Day belong to 7-days holiday.

Fig. 5 shows daily extremums of electricity demand in May Day 3-days holiday from 2015 to 2017. The day with a green circle is the legal holiday. It can be seen that daily extremums of electricity demand follow their own laws during holidays. For daily maximum of electricity demand, there are three tendencies, including uptrend (2015), downtrend (2017) and ``V"-shaped trend (2016). For daily minimum of electricity demand, there are two tendencies, including downtrend (2016 and 2017) and ``V"-shaped trend (2015). The events above, which also can be seen in other 3-days holidays, are mainly caused by the position of legal holiday in 3-days holiday.

Fig. 6 shows daily extremums of electricity demand in Spring Day 7-days holiday from 2015 to 2017. Because of the legal holiday in a fixed position during 7-days holiday, the tendencies for daily extremums of electricity demands are relative stables. Daily maximum of electricity demand shows a trend of "first decreased and then rose" while daily minimum of electricity demand has a wave-like increasing tendency. This kind of characteristic also can be seen in National Day.

3. The proposed methodology

Due to Section 2, it is beneficial for law mining to decompose the prediction for load profile into the predictions for proportional curve and daily extremum of electricity demand in holiday STLF problem.

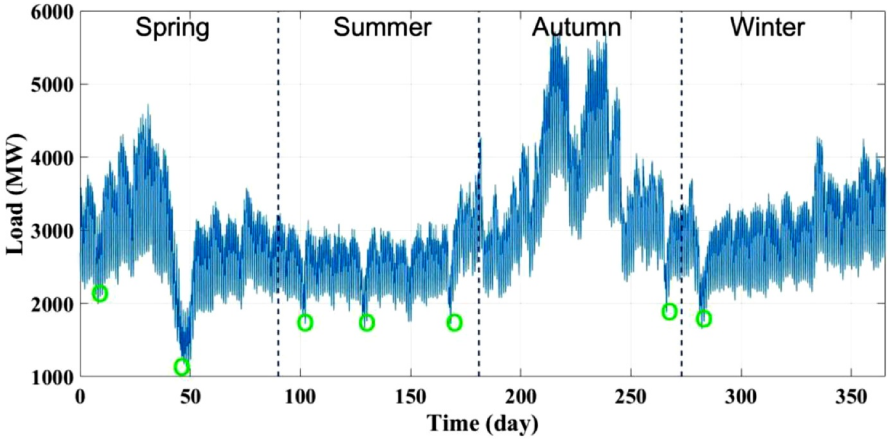

Fig. 1. Hourly electricity demand time series of provincial capital city in 2015.

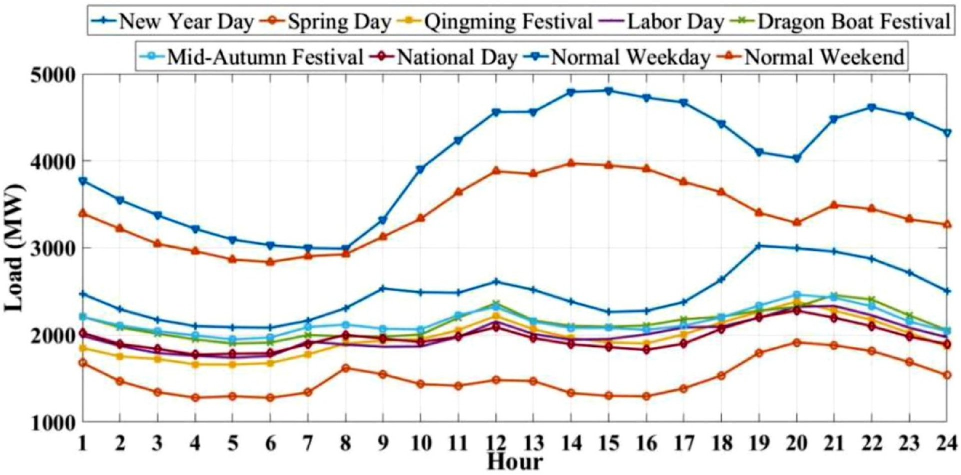

Fig. 2. Load profiles of New Year Day (1st Jan), Spring Day (19th Feb), Qingming Festival (5th Apr), May Day (1st May), Dragon Boat Festival (20th Jun), Mid-Autumn Festival (27th Sep), National Day (1st Oct), Normal Monday (4th Aug) and Normal Saturday (8th Aug).

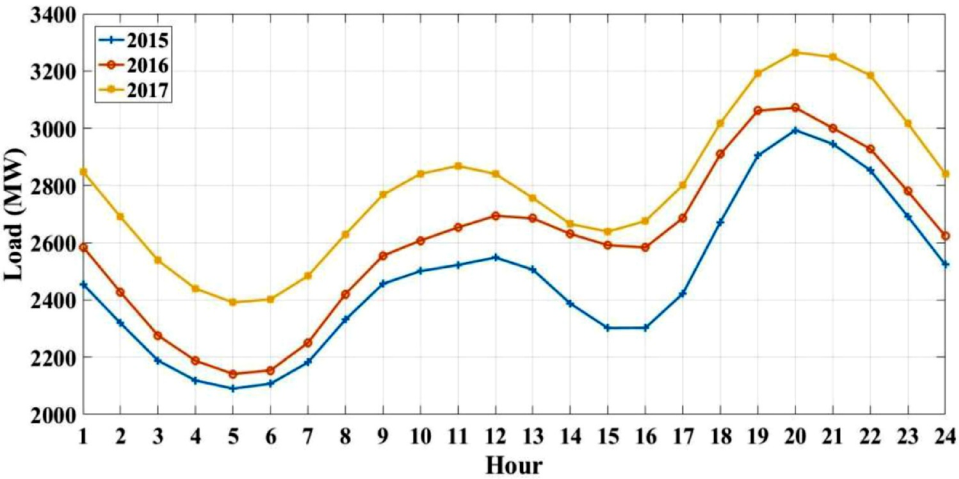

Fig. 3. Load profiles of new year day from 2015 to 2017.

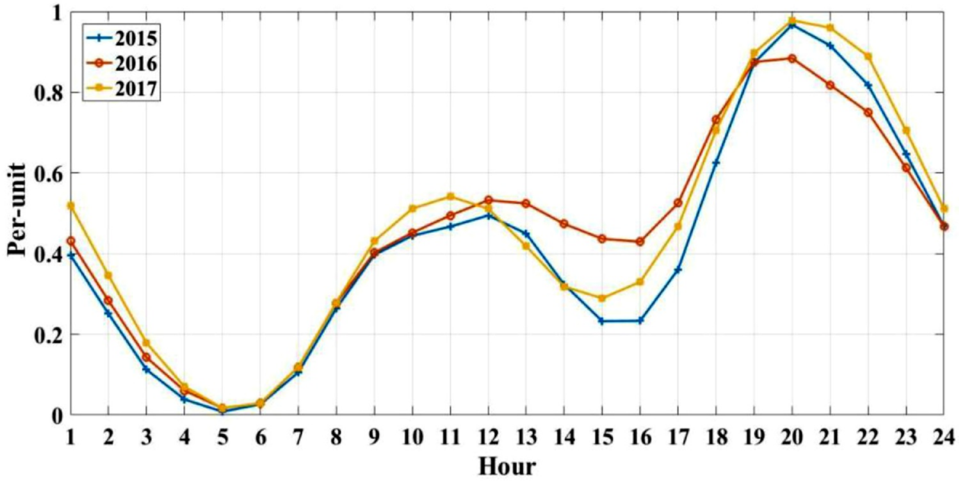

Fig. 4. Proportional curves of new year day from 2015 to 2017.

Thus, this paper presents a hybrid prediction model based on pattern sequence-based matching method and XGBoost. Its process is shown in Fig. 7.

3.1. *Prediction for proportional curve*

3.1.1. *Pattern sequence-based matching method*

Holiday proportional curve is mainly influenced by load composition, local customs and life style, which are the non-quantifiable factors.

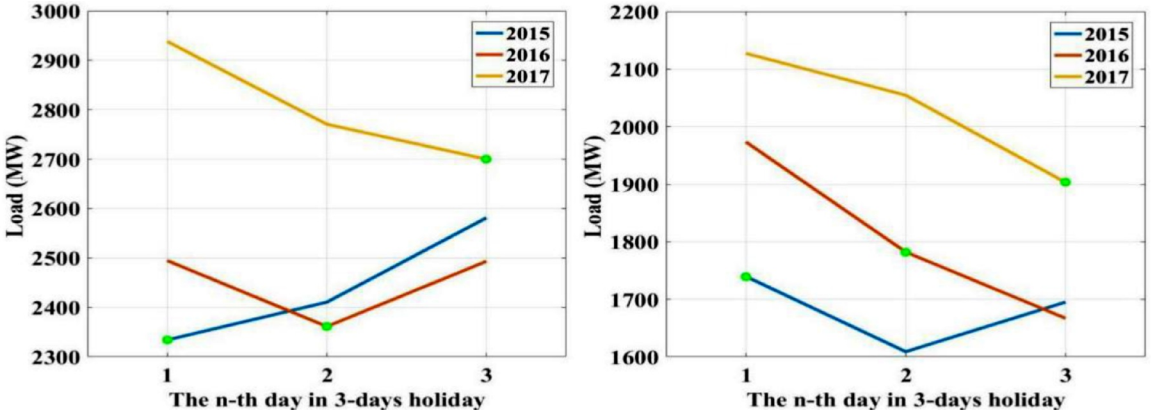

Fig. 5. Daily extremums of electricity demand in May Day 3-days holiday from 2015 to 2017 (left: **daily maximum**; right: **daily minimum**).

It is hard for the methods trained by load and meteorology data to predict the proportional curve. For solving this problem, pattern sequence-based matching method is adopted [21,22]. Its core idea is that the future trend of one time series can be inferred from the historical trend under the similar conditions. Thus, the proportional curves in two consecutive days can be seen as one pattern. Load regularity in the first day will have a significant impact on the load regularity in the second day with probabilities. Because of the non-quantifiable holiday factors contained in the proportional curve, day-ahead proportional curve can be suitable for predicting the proportional curve of the predicted holiday. So this method is built by the following assumption:

Proportional curve is defined as $\boldsymbol{p_h} = [p_{h,1}, \ldots, p_{h,\mathrm{T}}]$. T represents the total number of discrete time steps in one day. $[\boldsymbol{p}_{h\text{-}1}, \boldsymbol{p}_h]$ and $[\boldsymbol{p}_{h^*-1}, \boldsymbol{p}_{h^*}]$ represent two different proportional curves of two consecutive days, respectively. h and h^* represent different holidays. $h_{-}1$ and $h^* - 1$ represent the days ahead of h and h^*, respectively. If the known sequence $\boldsymbol{p}_{h-1}$ is similar to the historical sequence $\boldsymbol{p}_{h^*-1}$, then the predicted sequence $\boldsymbol{p}_h$ is similar to the historical sequence $\boldsymbol{p}_{h^*}$.

Fig. 8(a) shows day-ahead proportional curve $\boldsymbol{p}_{h-1}$ of the predicted holiday and similar historical day-head proportional curves $\boldsymbol{p}_{h^*-1}$. Fig. 8(b) shows the proportional curve $\boldsymbol{p}_h$ of the predicted holiday and the proportional curves $\boldsymbol{p}_{h^*}$ behind $\boldsymbol{p}_{h^*-1}$. It can be found that $\boldsymbol{p}_{h^*}$ are in the neighborhood of $\boldsymbol{p}_h$ for the predicted holiday, which proves the feasibility of pattern sequence-based matching method.

3.1.2. *Prediction algorithm*

Because proportion curves are normalized in interval of [0, 1] and their turning points are at the similar time, the Euclidean distance is used for measuring the similarity between two proportion curves. There exist the predicted holiday h and the historical holidays h^*. According to pattern sequence-based matching method, the similarity S between day-ahead proportion curve of the predicted holiday $\boldsymbol{p}_{h-1}$ and the historical one $\boldsymbol{p}_{h^*-1}$ is formulated as follows:

$$S(\boldsymbol{p}_{h-1}, \boldsymbol{p}_{h*-1}) = \sqrt{\sum_{t=1}^{T} (p_{h-1,t} - p_{h*-1,t})^2} \tag{2}$$

The smaller $S(\boldsymbol{p}_{h-1}, \boldsymbol{p}_{h^*-1})$ is, the smaller the difference between $\boldsymbol{p}_{h-1}$ and $\boldsymbol{p}_{h^*-1}$ is. Then, the historical holidays h^* is sorted by the similarities $S(\boldsymbol{p}_{h-1}, \boldsymbol{p}_{h^*-1})$ and the history similar date set $\boldsymbol{H}$ is selected from the most similar historical holidays:

$$\boldsymbol{H}=[h_1, \ldots, h_M] \tag{3}$$

where: h_M represents the historical holiday selected from h^*; M represents the number of the selected historical holidays.

So, the predicted proportional curve $\hat{\boldsymbol{p}}_h$ is obtained from the mean of the proportion curves in the history similar date set $\boldsymbol{H}$:

$$\hat{\boldsymbol{p}}_h = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{p}_{h_m} \tag{4}$$

It can be found that M has a significant impact on the prediction for proportional curve. Thus, the optimal value of M is determined by cross validation. During cross validation, the training data set $\boldsymbol{H}$ is divided into c subsets. In the each validation, one distinct subset is selected for validating the model trained by the remaining c-1 subsets. This procedure is repeated c times. The final estimation is obtained from the mean of the c validation results. Mean squared error (MSE) is selected as the model performance evaluation index:

$$\mathrm{MSE}_m = \frac{1}{c} \sum_{h \in H} (\hat{\boldsymbol{p}}_h - \boldsymbol{p}_h)^2 \{M = m\} \tag{5}$$

Then, the optimal value of M is the one that minimizes MSE:

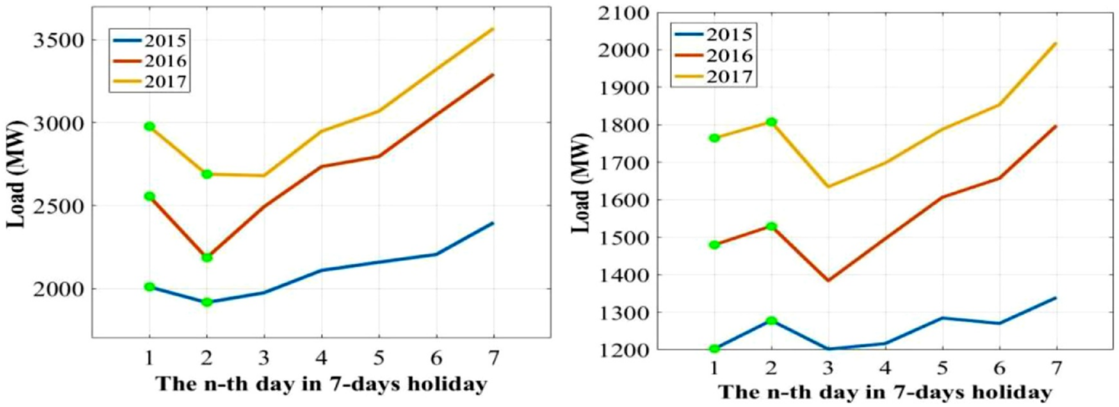

Fig. 6. Daily extremums of electricity demand in Spring Day 7-days holiday from 2015 to 2017. (left: **daily maximum**; right: **daily minimum**).

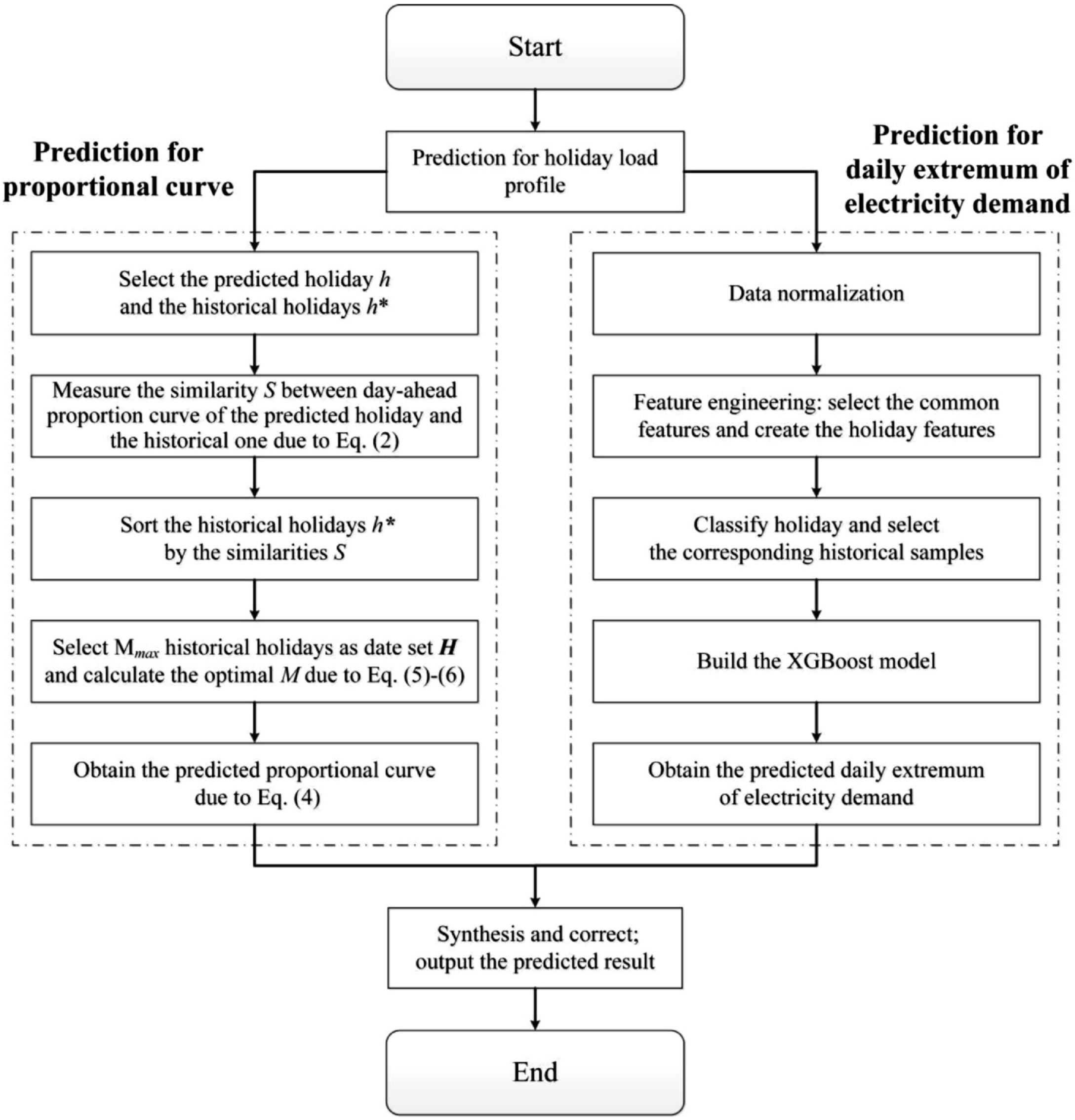
Start
Prediction for holiday load profile
Prediction for proportional curve
Prediction for daily extremum of electricity demand
Select the predicted holiday h and the historical holidays h*
Measure the similarity S between day-ahead proportion curve of the predicted holiday and the historical one due to Eq. (2)
Sort the historical holidays h* by the similarities S
Select M_max historical holidays as date set H and calculate the optimal M due to Eq. (5)-(6)
Obtain the predicted proportional curve due to Eq. (4)
Data normalization
Feature engineering: select the common features and create the holiday features
Classify holiday and select the corresponding historical samples
Build the XGBoost model
Obtain the predicted daily extremum of electricity demand
Synthesis and correct; output the predicted result
End

Fig. 7. The structure of the proposed methodology.

$$M = \arg\min\{MSE_m\}(m = 1, \ldots, M_{\max}) \tag{6}$$

where $M_{\max}$ is the setting value.

3.2. Prediction for daily extremum of electricity demand

Due to Section 2.3, daily extremum of electricity demand in holiday is mainly influenced by holiday effect, apart from climatic effect, load trend effect and temporal effect. Moreover, the holiday effects on daily extremum of electricity demand vary among holiday types, as shown in Figs. 5 and 6. Thus, the prediction for daily extremum of electricity demand is summarized in the following steps:

1) Data normalization;
2) Feature engineering: select the common features and create the holiday features;
3) Classify holiday by holiday length; select the historical samples with the same holiday type;
4) Build the XGBoost model trained by the selected historical samples;
5) Output daily maximum $\hat{P}_{h,\max}$ and minimum $\hat{P}_{h,\min}$ of electricity demand for the predicted holiday.

3.2.1. Data normalization

Considering the difference among data magnitudes, it is essential to normalize the data set, which can strengthen the model performance. Min-max normalization is used in mapping data between zero and one, described by Eq. (7):

$$\tilde{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{7}$$

where $\tilde{x}$ and x are the normalized and original value of indicator, respectively; $x_{\max}$ and $x_{\min}$ are the maximum and minimum value of the indicator, respectively.

3.2.2. Feature engineering

Feature engineering is vital to STLF, which directly affects the predictive result [23]. In this paper, we select features primarily according to the analysis of holiday load characteristics. Table 1 shows the features required for predicting holiday daily extremum of electricity demand. The inputs contains load vector, temperature vector, temporal vector and holiday vector. Input 1–3 represent the trend characteristics of load sequence. Input 5–8 can capture the temperature effects on load sequence during the same period. Input 9 and 10 represent the temporal characteristics of holiday load, which has a significant impact on holiday load base. Input 4 and Input 11–13 is new

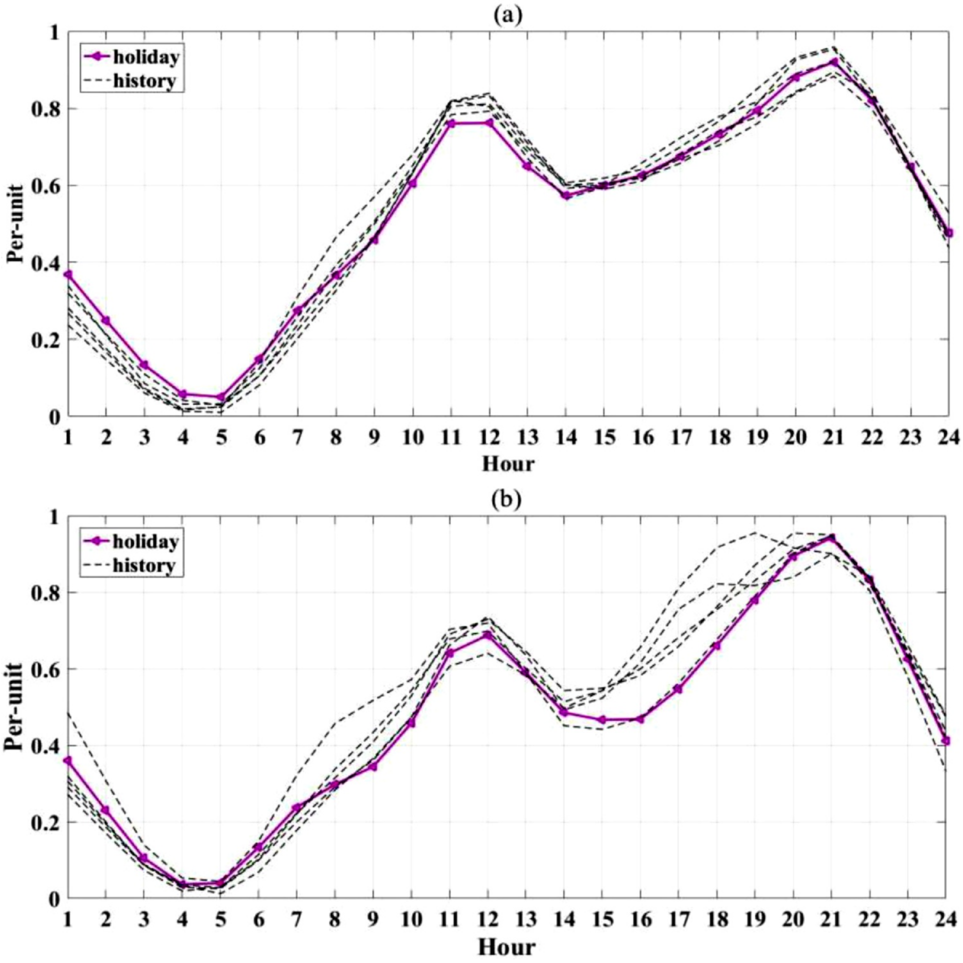

Fig. 8. The holiday demo for pattern sequence-based matching method.

feature created for holiday prediction. Input 4 represents the upper limit of holiday load. Input 11 classifies the holiday type by holiday length. In this paper, there exists two holiday types due to Section 2.3. Input 12 and 13 can capture the inflection points of holiday load, which determines the trend of holiday load. The above 13 inputs are expected to recognize the dynamic characteristics of holiday load.

3.2.3. *The modeling of XGBoost*

There are many matured predictive methods, such as SVM, NN and deep learning. However, these methods may not be suitable to the prediction for daily extreme of electricity demand in holiday. SVM has the advantages of high precision and strong generalization ability, but not robust to outliers. Although NN has strong ability of self-learning and non-linear expression, it has the shortcomings of the parameters of network, the uncertain unit number of the hidden layer and easily trapped to the local minimal, etc. The good performance of deep learning only lies in the massive high-dimensional data.

In recent years, a new ensemble learning algorithm XGBoost is proposed [24], which adopts decision tree as the base learner. It is a supervised algorithm that stacks all base learners into a strong learner. The predictive result by XGBoost is equal to the sum of all base learners. Inspired by the above idea, given the input *xi* normalized by Eq. (7), the final predictive value $\hat{y}_i$ is formulated as follows:

Table 1
Features required for predicting holiday daily extremum of electricity demand.

Input	Type	Name	Value / Unit
1	Relative to the previous holiday	Daily extremum of electricity demand	MW
2		Daily extremum of electricity demand before 24 h	MW
3		Daily extremum of electricity demand before 168 h	MW
4		The last daily extremum of electricity demand before holiday	MW
5		Temperature extremum	˚F
6		Temperature extremum before 24 h	˚F
7		Temperature extremum before 168 h	˚F
8	Relative to the predicted holiday	Temperature extremum	˚F
9		Season	1 - spring; 2 - summer; 3 - autumn; 4 - winter;
10		Day type	1–7 - Mon.~Sun.
11		Holiday type	1 - short; 0 - long
12		Holiday position	1–3 / 1–7
13		Legal holiday	1 - yes; 0 - no

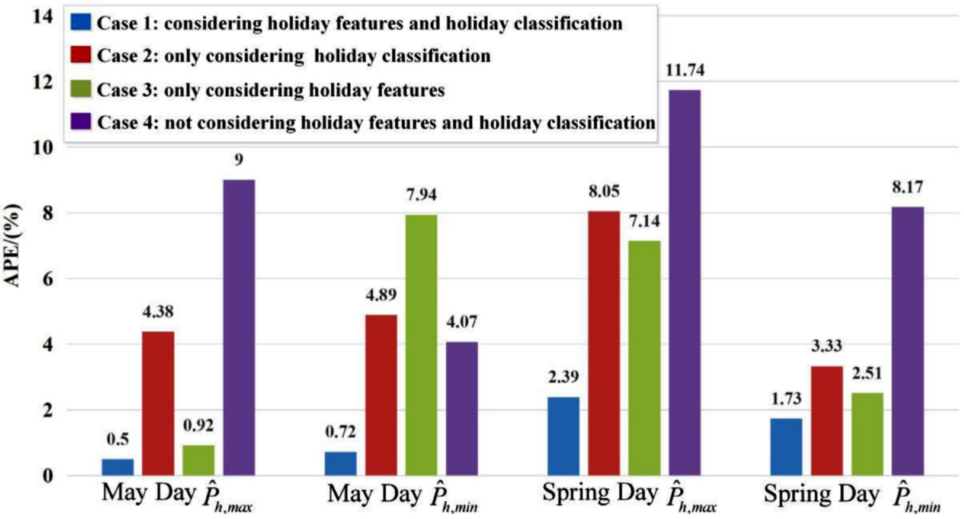

Fig. 9. Comparison of forecast errors for cases using and not using the holiday features and holiday classification.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{8}$$

where f_k represents the kth decision tree; K is the number of decision tree; F is the space that contains a set of decision trees.

In the process of regression, the object function of XGBoost is defined as:

$$Obj = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{9}$$

where $l(y_i, \hat{y}_i)$means the loss function, measuring the difference between the prediction$\hat{y}_i$ and the target y_i; Ω expresses the regularization term, measuring the model complexity; n indicates the number of target y_i.

The expression for Ω is as Eq. (10):

$$\Omega(f) = \gamma N_{node} + \frac{1}{2}\lambda \sum_{j=1}^{N_{node}} \omega_j^2 \tag{10}$$

where N_{node} indicates the number of leaf node in a decision tree; ω_j is the score of the jth leaf node; γ and λ represent the penalty factors.

Because of the additive model in Eq. (8), the forward stage wise algorithm is used in simplifying the model complexity. In each cycle when adding a decision tree, the model needs to learn the structure of the new function for fitting the last predicted residuals. In the zth cycle, the predictive value of xi: $\hat{y}_i^z = \hat{y}_{i-1}^{z-1} + f_z(x_i)$. Eq. (9) can be rewritten as shown below:

$$Obj^{(z)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^z) + \sum_{k=1}^{K} \Omega(f_k) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{z-1} + f_z(x_i)) + \Omega(f_z) \tag{11}$$

The greedy algorithm is adopted for building the decision tree based on the above object function. Then, a full model of XGBoost is completed by adding continuously decision trees.

3.3. Prediction for holiday load profile

Finally, the predicted holiday load profile $\hat{\boldsymbol{P}}_h$=[$\hat{P}_{h,1}$, ...,$\hat{P}_{h,\mathrm{T}}$] is synthesized from the prediction for proportional curve and daily extremum of electricity demand, as Eq. (12).

$$\hat{\boldsymbol{P}}_h = \hat{\boldsymbol{p}}_h \times (\hat{P}_{h,\max} - \hat{P}_{h,\min}) + \hat{P}_{h,\min} \tag{12}$$

In order to perfectly curve link together the predictive load profile and day-ahead load profile at 0:00 / 24:00, the segment correction between 0:00 and the moment of daily minimum $\hat{P}_{h,\min}$ of electricity demand is formulated as follows:

$$\hat{P}_{h,t} = \hat{P}_{h,\min} + (P_{h-1,\mathrm{T}} - \hat{P}_{h,\min}) \times \frac{t_{\min} - t}{t_{\min}}, t = 1, ..., t_{\min} \tag{13}$$

where $t_{\min}$ is the total of time step between 0:00 and the moment of daily minimum $\hat{P}_{h,\min}$ of electricity demand; $P_{h\text{-}1,T}$ represents the actual load at 24:00 in the day-ahead load profile.

4. Case study

To validate the proposed methodology, a typical case study on one provincial capital city in eastern China is carried out. Load data is selected with the sample time of 1 h. The training set is from year 2015 to 2017, and the testing set is in year 2018. For evaluation, the forecasting error is used with the following expressions:

$$\mathrm{APE} = \frac{|\hat{P} - P|}{P} \times 100\% \tag{14}$$

where $\hat{P}$ is the predicted value and P is the actual value.

4.1. The prediction analysis for daily extremum of electricity demand

May day (1st May) and Spring Day (16th Feb) in 2018 are used for verification in the prediction analysis for daily extremum of electricity demand. In Fig. 9, a comparison of forecasting errors for cases considering and not considering holiday features and holiday classification is conducted using XGBoost. It can be seen that the forecast errors are reduced by considering holiday feature created and holiday classification. The average errors of Case 1–4 are 1.34%, 5.16%, 4.63% and 8.25%, respectively. It demonstrates the importance of holiday features created and the need for holiday classification.

Moreover, the forecast errors by XGBoost are computed, compared with other algorithms, as shown in Fig. 10. This experiment is conducted considering holiday features and holiday classification. The average errors are 1.34%, 4.93%, 2.97%, and 2.83% for XGBoost, SVM, BPNN (back propagation neural network), and RF (random forest), respectively. This means that XGBoost outperforms other algorithms in prediction. Despite some algorithms could perform better on certain cases, such as BPNN in Spring Day $\hat{P}_{h,\min}$, XGBoost is more stable and reliable to predict the daily extremum of electricity demand.

4.2. The prediction analysis for the proposed methodology

In this experiment, the proposed methodology is used in holiday STLF of May Day 3-days holiday and Spring Day 7-days holiday in 2018. Figs. 11 and 12 show their diagrams concerning the predicted

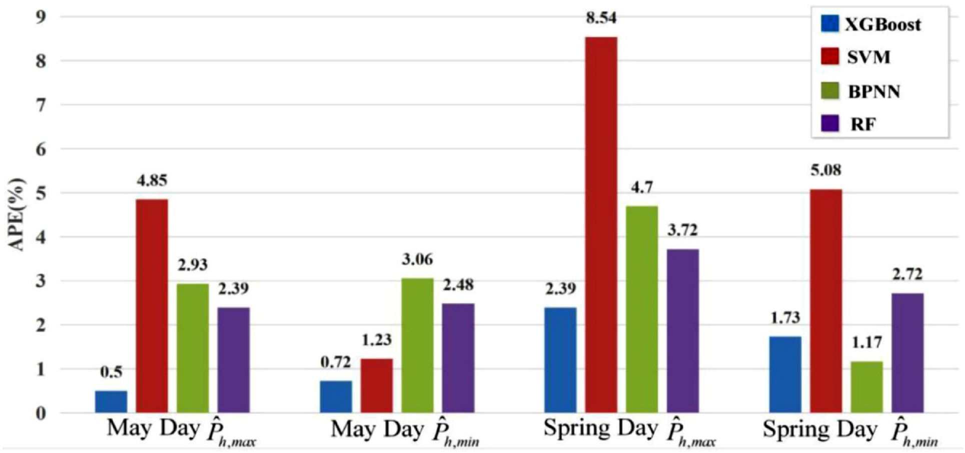

Fig. 10. Comparison of forecast errors by different algorithms.

and actual load curves. Their predictive results are also seen in Tables 2 and 3.

From Figs. 11 and 12, the proposed methodology can grasp the main holiday load characteristics with a relatively high accuracy. The average errors of 24 h load forecast errors for May Day 3-day holiday and Spring Day 7-day holiday are in a very small range, 3.28% and 2.78%, respectively. Most errors of daily extremum of electricity demand are in 2.5%. Although some days exist little gaps between the predicted and actual load curves at some points, the main reason is that too high holiday effects are difficult to handle with. Take 30th April and 17th Feb as examples. The midday sharp reduction rarely happens in the history. Despite the prediction for daily extremum of electricity demand is satisfactory, it is hard to find a historical similar proportional curve for matching, which leads to the larger forecast errors. Another example is 15th Feb. Despite the predicted and actual load curves are similar in load shape, the inaccurate prediction for daily maximum of electricity demand also influences the accuracy. But overall, the proposed methodology are acceptable for holiday STLF.

Moreover, the forecast errors of all the holidays in 2018 by the proposed methodology are computed, compared with other algorithms. Table 4 shows the holiday comparison of 24 h load forecast errors by different algorithms in 2018. Through comparison results can be drawn the following conclusion:

1) The proposed methodology gives much better forecasting accuracy with an average error 2.98%. Compared with other methods, about 1/4 to 1/2 of forecasting error is reduced.
2) Because of weak robustness, SVM can not play to its advantages in holiday STLF. Although BPNN performs better than the proposed methodology in some cases, it loses its stability. RF has the advantages of high prediction accuracy, strong robustness and few parameters, avoiding over-fitting effectively. However, RF is originally designed for classification problem, so it can not take its advantages in STLF problem. Moreover, these algorithms are modeled by holiday load profile, whose characteristics are complex and difficult to mine directly. That is why the accuracy can not be improved by these algorithms.
3) The proposed methodology divides holiday STLF problem into two prediction problems. The prediction for proportional curve can enhance the shape similarity of holiday load curve and overcome the problem of small observation set in holiday STLF problem. The prediction for daily extremum of electricity demand takes many factor effects headed holiday features and holiday classification into account, which contributes to the accuracy improvement of holiday STLF.

To verify the applicability of the proposed methodology, it has also applied to the other surrounding cities in the same province. Table 5 provides the mean of APE (MAPE) of holiday STLF using different algorithms on each city in 2018. From Table 5, we can observe that the proposed methodology outperforms the other algorithms in most cities. The averaged improvement of forecasting accuracy using the proposed methodology is between 0.71% and 2.98%. Hence, our methodology is robust and adaptive to different cities, which ensures its high precision in holiday STLF.

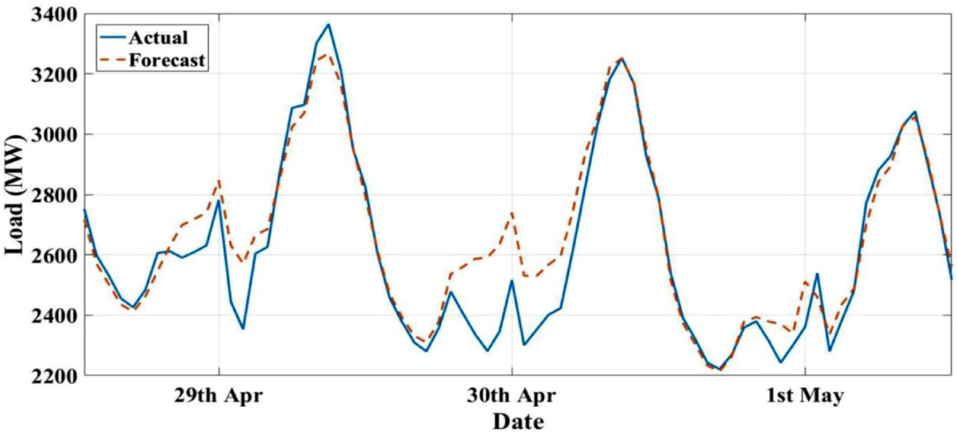

Fig. 11. Holiday load forecasting for May Day 3-days holiday in 2018.

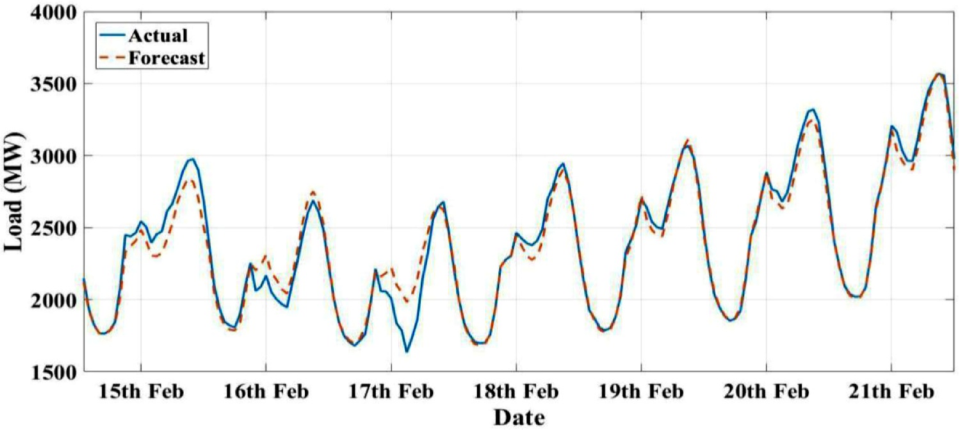

Fig. 12. Holiday load forecasting for Spring Day 7-days holiday in 2018.

Table 2
The forecast results for May Day 3-days holiday in 2018.

Date	Error of daily maximum of electricity demand	Error of daily minimum of electricity demand	Average error of 24 h load forecast	Maximum error of 24 h load forecast
4.29	2.10%	1.88%	2.71%	8.43%
4.30	1.60%	1.10%	4.96%	12.7%
5.1	0.50%	0.72%	2.16%	4.54%
Average	1.40%	1.23%	3.28%	8.56%

Table 3
The forecast results for Spring Day 7-days holiday in 2018.

Date	Error of daily maximum of electricity demand	Error of daily minimum of electricity demand	Average error of 24 h load forecast	Maximum error of 24 h load forecast
2.15	5.71%	0.93%	2.47%	6.88%
2.16	2.39%	1.73%	3.10%	5.96%
2.17	0.47%	0.61%	5.55%	25.3%
2.18	1.84%	1.11%	2.27%	5.34%
2.19	1.90%	0.80%	2.08%	3.18%
2.20	2.13%	1.14%	2.43%	3.55%
2.21	0.88%	1.05%	1.55%	2.74%
Average	2.19%	1.05%	2.78%	7.56%

5. Conclusion

In this paper, a novel hybrid prediction model based on pattern sequence-based matching method and XGBoost is proposed for holiday STLF problem. Two essential properties of holiday load characteristics are introduced, namely proportional curve and daily extremum of electricity demand. The properties of each holiday load profile can be uniquely identified by them. Then, the proposed methodology divides holiday STLF problem into two independent prediction tasks, which contributes to the law mining of holiday load characteristics. We mainly compare our methodology with many algorithms on one provincial capital city to illustrate its feasibility. The case studies show that the performance metrics are improved. From the comparative experimental analysis for predicting daily extremum of electricity demand, the improvement is due to the holiday feature created, holiday classification and the selected XGBoost model. From the comparative experimental analysis for predicting holiday load profile, the improvement is due to the mechanisms of task decomposition and synthesis and pattern sequence-based matching method.

Table 5
MAPE of holiday STLF using different algorithms on each city in 2018.

Target city	The proposed methodology	BPNN	Similar day method	SVM	RF
Capital city	2.98%	4.24%	6.45%	6.67%	3.96%
City #01	3.82%	4.53%	5.00%	6.59%	4.78%
City #02	4.38%	5.76%	4.78%	7.10%	4.29%
City #03	2.73%	2.84%	3.60%	5.03%	3.53%
City #04	3.04%	3.86%	4.43%	6.46%	3.94%
Average	3.39%	4.25%	4.85%	6.37%	4.10%

CRediT authorship contribution statement

Kedong Zhu: Conceptualization, Methodology, Software, Writing -

Table 4
Holiday comparison of 24 h load forecast errors by different algorithms in 2018.

Holiday type	Holiday Name	The proposed methodology	BPNN	Similar day method	SVM	RF
3-days holiday	New Year Day	2.84%	4.27%	6.51%	7.79%	3.64%
	Qingming Festival	3.19%	2.79%	6.62%	6.11%	3.71%
	May Day	3.28%	4.81%	5.90%	5.28%	3.81%
	Dragon Boat Festival	2.70%	4.91%	5.16%	6.70%	3.51%
	Mid-Autumn Festival	2.78%	3.28%	5.91%	6.78%	3.75%
7-days holiday	Spring Day	2.78%	5.12%	6.87%	7.26%	4.34%
	National Day	3.21%	3.83%	6.97%	6.37%	4.17%
Aeverage	–	2.98%	4.24%	6.45%	6.67%	3.96%

original draft, Writing - review & editing. **Jian Geng:** Supervision, Validation, Investigation. **Ke Wang:** Validation, Data curation, Resources.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by State Grid Corporation of China Project “Research on Bulk Power Grid `Source-Grid–Load-Storage' Coordinated Control and Intra-day Control for Flexible Loads” (Grant No. SGJS0000DKJS1900257) and National Natural Science Foundation of China (Grant No. 51807181).

References

[1] Z. Shao, F. Gao, Q. Zhang, et al., Multivariate statistical and similarity measure based semiparametric modeling of the probability distribution: a novel approach to the case study of mid-long term electricity consumption forecasting in China, Appl. Energy 156 (Oct 2015) 502–518.

[2] L. Quilumba, W. Lee, H. Huang, et al., Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities, IEEE Trans. Smart Grid 6 (2) (Mar 2015) 911–918.

[3] Y. Li, D. Han, Z. Yan, Long-term system load forecasting based on data-driven linear clustering method, J. Mod. Power Syst. Clean Energy 6 (May 2017) 306–316.

[4] M. Rana, I. Koprinska, Forecasting electricity load with advanced wavelet neural networks, Neurocomputing 182 (Mar 2016) 118–132.

[5] E. Ceperic, V. Ceperic, A. Baric, A strategy for short-term load forecasting by support vector regression machines, IEEE Trans. Power Syst. 28 (4) (Nov 2013) 4356–4364.

[6] J. Nobrega, A. Oliveira, Kalman filter-based method for online sequential extreme learning machine for regression problems, Eng. Appl. Artif. Intell. 44 (Sep 2015) 101–110.

[7] D. Chaturvedi, A. Sinha, O. Malik, Short term load forecast using fuzzy logic and wavelet transform integrated generalized neural network, Electr. Power Energy Syst. 67 (May 2015) 230–237.

[8] R. Hyndman, K. Ord, R. Snyder, et al., Forecasting time series with multiple seasonal patterns, Electr. Power Energy Syst. 191 (1) (Nov 2008) 207–222.

[9] M. Kim, Modeling special-day effects for forecasting intraday electricity demand, Eur. J. Oper. Res. 230 (1) (Oct 2013) 170–180.

[10] Y. Wi, S. Joo, K. Song, Holiday load forecasting using fuzzy polynomial regression with weather feature selection and adjustment, IEEE Trans. Power Syst. 27 (2) (May 2012) 596–603.

[11] A. Ebrahimi, A. Moshari, Holidays short-term load forecasting using fuzzy improved similar day method, Int. Trans. Electr. Energy Syst. 23 (8) (Nov 2013) 1254–1271.

[12] S. Arora, J. Taylor, Short-Term forecasting of anomalous load using rule-based triple seasonal methods, IEEE Trans. Power Syst. 28 (3) (Aug 2013) 3235–3242.

[13] S. Arora, J. Taylor, Rule-based autoregressive moving average models for forecasting load on special days a case study for France, Eur. J. Oper. Res. 266 (3) (Sep 2017) 259–268.

[14] B. Li, J. Huang, Y. Wu, et al., Holiday short-term load forecasting based on fractal characteristic modified meteorological similar day, Power Syst. Technol. 41 (6) (Jun 2017) 1949–1955.

[15] H. Lin, J. Liu, Z. Feng, et al., Short-term load forecasting for holidays based on the similar days'load modification, Power Syst. Prot. Control 38 (7) (Apr 2010) 47–51.

[16] N. Fidalgo, J. Lopes, Load forecasting performance enhancement when facing anomalous events, IEEE Trans. Power Syst. 20 (1) (2005) 408–415.

[17] L. Feng, J. Qiu, Short-term load forecasting for anomalous days based on fuzzy multi-objective genetic optimization algorithm, Proc. CSEE 25 (10) (May 2005) 29–34.

[18] Z. Florian, Modeling public holidays in load forecasting: a German case study, J. Mod. Power Syst. Clean Energy 6 (2) (Feb 2018) 191–207.

[19] P. Zeng, C. Sheng, M. Jin, A learning framework based on weighted knowledge transfer for holiday load forecasting, J. Mod. Power Syst. Clean Energys 7 (2) (Sep 2019) 329–339.

[20] G. Cerne, D. Dovzan, I. Skrjanc, Short-term load forecasting by separating daily profiles and using a single fuzzy model across the entire domain, IEEE Trans. Ind. Electron. 65 (9) (Sep 2018) 7406–7415.

[21] C. Jin, G. Pok, H. Park, et al., Improved pattern sequence-based forecasting method for electricity load, IEEJ Trans. Electr. Electron. Eng. 9 (Sep 2014) 670–674.

[22] M. Francisco, A. Troncoso, J. Riquelme, et al., Energy time series forecasting based on pattern sequence similarity, IEEE Trans. Knowl. Data Eng. 23 (8) (Aug 2011) 1230–1243.

[23] Z. Hu, Y. Bao, T. Xiong, et al., Hybrid filter-wrapper feature selection for short-term load forecasting, Eng. Appl. Artif. Intell. 40 (Apr 2015) 17–27.

[24] S. Taieb, R. Hyndman, A gradient boosting approach to the Kaggle load forecasting competition, Int. J. Forecast. 30 (2) (Apr 2014) 382–394.