Short-term Power Load Forecasting Based on Clustering and XGBoost Method

Yahui Liu, Huan Luo
School of Information Management
Beijing Information Science and Technology University
Beijing, China
yahui.liu@126.com

Bing Zhao
Metrology Department
China Electric Power Research Institute
Beijing, China

Xiaoyong Zhao, Zongda Han
School of Information Management
Beijing Information Science and Technology University
Beijing, China
zhaoxiaoyongtisbistu.edu.cn

Abstract- According to the problems of high computattonal cost and over-fitting in traditional forecasting methods, a short-term power load forcasting method is put forward based on combining clustering with xgboost (eXtreme Gradient Boosting) algorithm. The method mainly does research on correlation between influence factors and load forecasting results. Firstly, Features extracted from original datum and mlssmg values are filled during preprocessmg stage. Secondly, the changing trend of load is divided into four classifications by K-means algorithm. Meanwhile, classification rules are set up between temperature and category. Finally, xgboost regression model is established for different classifications separately. Furthermore, forecasting load is calculated according to scheduled date. Experimental results indicate the method can to some extent predict the daily load accurately,

Keywords-clustering; decision tree; xgboost; power load forecasting; multiple classification

I. INTRODUCTION

It is necessary to carry out power load forecasting in order to ensure the rationality of production management and draw up a plan. Furthermore, it can improve the security, reliability and stability of power system by estimating electricity consumption in some slotl'! Traditional forecasting algorithms include time series method, regression analysis method, the trend analysis approach and exponential smoothing etc. Merits of time series method lie in short computing time, but high fluctuation of time series may have influence on accuracy. Regression analysis method applies to low-order model which is of good computational efficiency, but spurious regression may be caused by inaccuracy hypothesis of regression equation. In addition, feedforward neural network is of high computational cost and easy to be overfitting [2-4]. Accuracy of power load prediction is mainly affected by temperature, area, holiday, forecasting model and integrity of historical data etc.

II. RELATED WORK

Srivastava A K et al. [5] provide an overview of short-term power load prediction. Furthermore, Kasim Zor et al.[6] divide forecasting algorithms into two classifications. One is traditional methods used in normal conditions such as linear regression, Box-Jenkins method, distribution-free regression procedure and so on. But there exist limits to changes of climate, society and economy. The other is methods related to artificial intelligence such as neural network, Support Vector Machine, neuro fuzzy inference system. Liu D et al[7]. make use of distributed short-term load forecasting to solve problems in large area. Furthermore, it does research on changes of load and weather, which have influence on forecasting results.

Silva et al.[8] conduct short-time load forecasting to power distribution substations by adaptive neuro fuzzy inference system. With expansion of power network capacity and rapid increase of data quantity, traditional forecasting algorithms exist lacks of high complexity and high time cost. Furthermore, they are hardly applicable to parallel processing.

Integrated learning algorithms mainly include xgboost (Extreme Gradient Boosting) , RF and GBRT etc. Merits of RF(Random Forest) are easy to realize, low computational complexity,but it tends to be overfitting under noisy condition. Although GBDT(Gradient Boosting Decision Tree) can be used to deal with multiple types of datum well, it is difficult to handle parallel data. As to xgboost, it is effective to prevent overfitting, reduce computational complex and it can be used in parallel processing. Wang[9]et al. make use of RF combined with temperature and wind velocity to predict load on the hadoop platform. Papadopoulos Spol et al. compare Root Mean Square Error(RMSE) and Mean Absolute Percentage Error (MAPE) aiming at following four methods: SARlMA, SARlMAX, RF and GBRT.

Short-term load forecasting and related analysis is carried out on the basis of consumption behavior, temperature and date. First of all, it uses K-means to divide preprocessing datum into different categories. Then, classification rules are established by decision tree. Finally, regression model is set up by xgboost so as to realize daily load forecasting function. In the paper, it includes four parts: power load forecasting procedures are introduced in section I; principal and algorithm are described in section II; implementation process is presented in section III; experimental results are analyzed in section IV.

978-1-5386-6565-7/18/$31.00 ©2018 IEEE

III. POWER LOAD FORECASTING PROCESS

Short-load forecasting procedures are described in figure I:

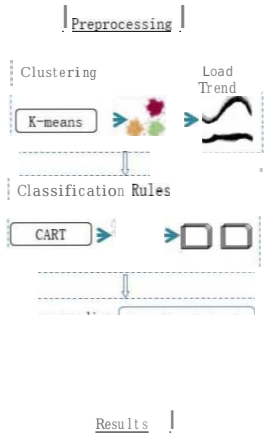


Figure 1. FORECASTING FLOW

Step1: at preprocessing stage, missing datum are filled by means of average values. Then, effective data are standardized and divided into groups by day.

Step2: classification rules divided by K-means are set up in the light of temperature, date and clustering kind. Feature extractions of data mainly involve holiday, the top temperature, average temperature, time of which are from 6 to 10 and from 12 to IS separately. The smallest Gini index is used as partition attribute in decision tree.

Step3: the forecasting model is put forward for each load classification. Furthermore, xgboost regression model is trained according to four different divided categories.

A. Load classification division

It is to seen daily load as a vector. Historical loads are divided into four kinds by K-means algorithm according to holiday, weekend etc. Furthermore, k points are chosen as initial centroids. Meanwhile, distances are calculated, which are between centroids and data points. Each datum point is distributed to the nearest centroid. What's more, centroids are updated by mean values of distance. Finally, distances between sample vectors and centroids belonged to are minimized. Samples distributed and clustering centroids chosen are iteration until to convergence.

Quadratic sum of distance within each cluster is shown in Eq. (I):

$$SSD = \sum_{l=1}^{k} \sum_{x \in C_i} \| x - \mu_i \|_2^2 \qquad \mu_i = \frac{1}{C_i} \sum_{x \in C_i} x \qquad (1)$$

Where μ_i is the mean vector of cluster Cr. In a cluster, the higher value of SSD is, the lower of sample similarity is.

Silhouette coefficient is used to measure the effect of clustering results and its range of value is within [-1,1]. Meanwhile, the higher silhouette coefficient is, the closer distance of similar samples is. Silhouette coefficient formula is in Eq. (2):

$$S = \frac{d_1 - d_2}{\max(d_1, d_2)}$$

As to each sample, d_2 is the mean distance which is between it and other homogeneous samples within a cluster, while d_1 is that between it and points out of cluster (III.

B. Classification rule establishment

Relationship is set up by CART, which is between decision trees and influence factors. Set $D(X_1, X_2, \cdots, X_n)$ refers to 24h load values in some day and Xi is a vector. Purity of set D is evaluated by Gini index, where Gini index is shown in Eq. (3):

$$\mathrm{Gini}(D) = 1 - \sum_{i=1}^{|y|} P_i^2 \qquad (3)$$

Pi represents quantity proportion of category i account for those of total samples, where |y| is the numbers of total samples. Establishment of decision tree is described as follows:

Four features are obtained from preprocessing data, which include mean temperature from 6 o'clock to 10 o'clock and from 12 to 15 separately, the maximum temperature and holiday. Then, it's to choose the minimum Gini index as the split attribute. At last, the split process is iterated until to approximate leaf nodes.

C. Power load prediction model

Xgboost model is shown in Eq. (4), where Γ is the space of regression tree and K is additive function [12-15].

$$\hat{y}_i = \sum_{k=1}^{k} f_k(x_i), f_k \in \Gamma \qquad (4)$$

Objective function is described in Eq. (5).

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \qquad (5)$$

Where $l(y_i, \hat{y}_i)$ is a loss function, $\Omega(f_k)$ is a regular term shown in Eq. (6), which is used to measure the model complexity. T is the numbers of leave nodes and γ, λ are weight and coefficient separately.

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} \omega_j^2$$

For the purpose of minimize objective function, it is necessary to adjust parameters. Results are shown in Eq. (7) by optimizing objective function.

$$Obj^{(t)} = \sum_{j=1}^{T} [G_j \omega_j + \frac{1}{2}(H_j + \lambda)\omega_j^2] + \gamma T \qquad (7)$$

Where $G_j = \sum_{i \in I_j} g_i$ $H_j = \sum_{i \in I_j} h_i$ Weight and coefficient among leaf nodes are independence each other. The optimal value of each leaf node and objective function are shown in Eq. (8):

$$\omega_j^* = -\frac{G_j}{H_j + \lambda}$$

$$Obj^* = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j + \lambda} + \gamma T \qquad (8)$$

As for each segmentation point, it is to choose the higher Gain value to split. Gain calculation is ShO\\11in Eq. (9).

$$Gain = \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} - \gamma$$

D. Experimental analysis

The experimental data are origin from the open data and it does research on commercial power in the paper. Trend of power load forecasting in some area within 1644 days is shown in Fig 2 by data preprocessing, where the unit is myriawatt.

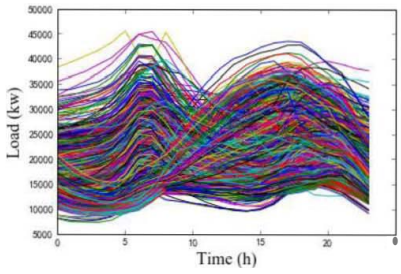


Figure 2. DAILY LOAD CURVE

- After data normalization, it makes use of k-means clustering algorithm to classify load data. X-axis represents 24h and Y-axis points out load data normalized shown in Fig.3.

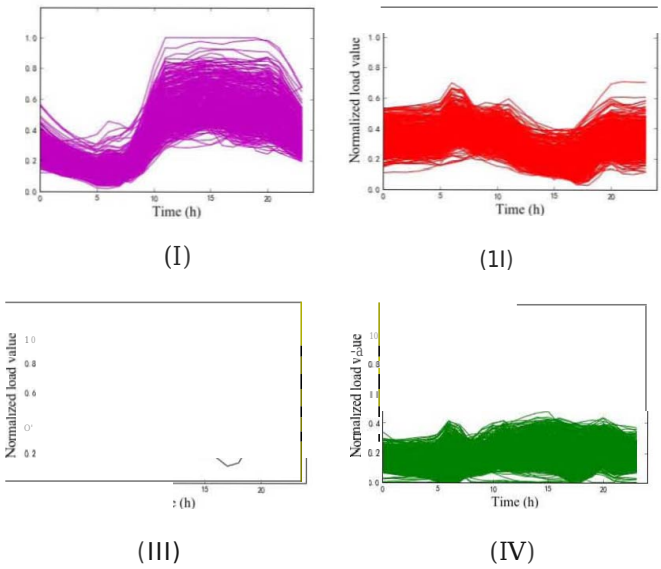


Figure 3. LOAD CLASSIFICATIONS

- Statistical data are divided into four types: I is weekend load; II and IV are mainly daily load; III is holiday load within 365 days. Relationship between a day and classification is shown in table.I.

TABLE I. RELATIONSHIP BEIWEEN A DAY AND CLASSIFICATION

Classification / Week	I	II	III	IV
Monday	25	8	13	6
Tuesday	19	12	13	7
Wednesday	20	13	13	6
Thursdav	27	9	10	7
Fridav	30	7	12	4
Saturday	40	2	7	4
Sunday	40	1	8	3

Computational complexity of multiple decision trees is reduced by feature selection and pruning method, which is also for the purpose of avoiding overfitting. While pruned, the depth of the most accurate tree is calculated by training and testing data set. Features mainly involve temperature and date in CART. In details, they include daily mean temperature, the minimum and maximum temperature, date category etc. Local details are shown in FigA. Accuracy of CART model is approximate 0.85.

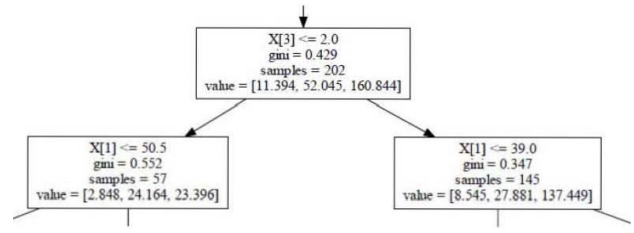


Figure 4. CART ESTABLISHMENT

- Forecasting dates are divided into corresponding classification by influence factors shown in table 2.

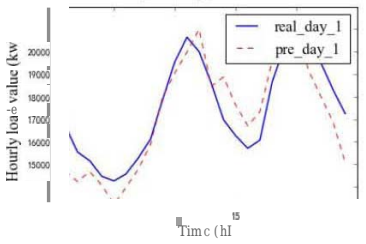
TABLE II. DATE AND CLASSIFICATION

Forecasting Day	Classification
Jan.1	I
Dec.12	III
Jan.21	IV
Jun.29	II

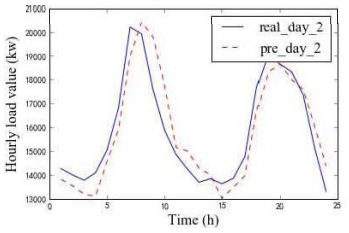
- Final forecastmg results are obtained by CART and xgboost model and error ε is computed ShO\\l1 in Eq. (10), where i E N and N is an integer within 24h. Experimental error is approximate 12%.

$$\varepsilon = \frac{1}{n}\sum_{i=1}^{n}\left|(\hat{y} - y)\right| / \frac{1}{n}\sum_{i=1}^{n} y \qquad (10)$$

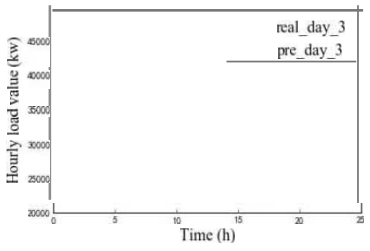
Comparison between true value and predicted value are shown in Fig.5.



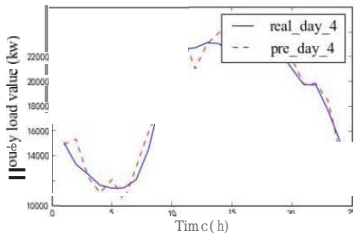
Jan.



Dec.12



Jan.21



Jun.29

Figure 5. COMPARE TRUE VALUE WITH FORECASTING VALUE

Horizontal ordinate represents 24h and Y-axis refers to the load value.

E. Conclusion

In order to solve the problems of overfitting and high computational complexity during the research of short-term load forecasting, it combines K-means clustering, CART, xgboost with temperature and date factors. Clustering is obtained according to the trend of daily load changes. Then, classification rules are set up by CART, which are combined temperature with special date factors etc. Finally, xgboost regression model is established to different categories. In addition, limited influence factors may have influence on the accuracy of experimental results. Experimental analyses indicate that the method can divide classification and forecast load value of appointed date accurately.

ACKNOWLEDGMENT

Science and Technology Planning Project of Beijing Municipal Commission of Education (KM201711232018).

REFERENCES

[1] Lai Zhengtian. Electric Power Big Data[M]. China Machine Press,2016,pp.96-100.

[2] Lang Kun. Research on Short-term Load Forecasting and Optimization of Economic Dispatch Decision-making of Electric Power System[D]. Dahan University of Technology, 2016, pp.5-8.

[3] Jin Xin, Li Longx:ian et al. Power Short-term Load Forecasting Based on Big Data and Optimization Neural Network[J]. Journal on Communications, 2016(sl), pp.36-42.

[4] Li Donghui, Yin Haiyan et al. An Annual Load Forecasting Model Based on Generalized Regression Neural Network with Multi-Swarm Fruit Fly Optimization Algorithm[J]. Power System Technology, 2018(2).

[5] Srivastava A K, Pandey A S, Singh D. Short-term Load Forecasting Methods: A Review[C]// International Conference on Emerging Trends in Electrical Electronics & Sustainable Energy Systems. IEEE, 2016.

[6] Zor K, Timur 0, Teke A, et al. A State-of-the-art Review of Artificial Intelligence Teclmiques for Short-term Electric Load Forecasting[C]// International Youth Conference on Energy. 2017, pp.I-7.

[7] Liu D, Zeng L, Li C, et al. A Distributed Short-Term Load Forecasting Method Based on Local Weather Information[J]. IEEE Systems Journal, 2016, pp.l-S.

[8] Silva I N D, Andrade L C M D. Efficient Neuro Fuzzy Model to Very Short-Term Load Forecasting[J]. IEEE Latin America Transactions, 2016, 14(2), pp.721-728.

[9] Wang Dewen, Sun Zhiwei. Big Data Analysis and Parallel Load Forecasting of Electric Power User Side[J]. Proceedings of the CSEE, 2015, 35(3):527-537.

[10] Papadopoulos S, Karakatsanis I. Short-term Electricity Load Forecasting Using Time Series and Ensemble Learning Methods[C]// Power and Energy Conference at illinois. IEEE, 2015, pp.1-6.

[11] Zhou Zhihua. Machine Learning[M]. Tsinghua University Press, 2016, pp.180-181.

[12] Chen T, Guestrin C. Xgboost , Reliable Large Scale Tree Boosting System[C]. Proceedings of the 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA. 2016, pp.13-17.

[13] Li Guangye. Short-Term Electricity Load Forecasting Based on the Xgboost Algorithm],l]. Smart Grid.2017, 07(4), pp.274-285.

[14] Zhang D, Qian L, Mao B, et al. A Data-Driven Design for Fault Detection of Wind Turbines Using Random Forests and XGBoost[J]. IEEE Access, 2018, PP(99),pp.1.

[15] Zhang P, Wu X, Wang X, et al. Short-term Load Forecasting Based on Big Data Technologies],l]. Csee Journal of Power & Energy Systems, 2015, 687-691(3), pp.1186-1192.