Abnormal Power Consumption Detection Based on Data-Driven

Jiang Jianfeng [1], Zhu Wenjun [1], Wang Xingang[1] and Zhang Chong[2*]

[1]Electric Power Research Institute, State Grid Shanghai Municipal Electric Power Company, Shanghai, China
[2]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China

Abstract. Based on high dimensional random matrix theory and machine learning algorithm, a method to detect abnormal power consumption behaviour of users is proposed. Firstly, the K-means clustering algorithm is used to divide the power loads into load types that obey specific distribution law or with random fluctuation. Then the linear eigenvalue statistics (LES) index can be used to detect the abnormal power consumption behaviour for the former such as unimodal load or bimodal load. And the difference between the actual and predicted value of regression model based on XGBoost algorithm can be used as the basis for judging abnormal power consumption behaviour of the latter. The method proposed in this paper is applicable to different types of loads and can implement a good discriminant effect.

1 Engineering background and issues

Electricity theft is one of the principal causes for the revenue losses of distribution utilities. The traditional abnormal identification of electricity consumption behaviour needs to invest a lot of manpower[1], financial resources and material resources, and can not achieve the desired effect.

These mechanisms may become very difficult due to dynamically varying and inadequate load flow information. Based on fine-grained data collected from networks, devices and consumers, distribution utilities are developing the analytical capabilities for improved detection of electricity theft. Popular data-driven methods include load profile analysis and loss calculation. Load profile methods identify suspected electricity thefts by classifying abnormal consumption patterns using data mining techniques. Domestic and foreign scholars are using power big data to analyze users' abnormal use the electrical behavior recognition has been studied. In reference [2], clustering algorithm is used to analyze the power load characteristics of each type of users and obtain characteristic curves. The deviation degree of load curve and characteristic curve is used to judge whether the user behavior is abnormal. In reference [3], an abnormal electricity consumption pattern detection model based on unsupervised learning is proposed, the model includes feature extraction, principal component analysis, mesh processing, calculation of local outlier factors and other modules. Reference [4] proposes the electric theft detection model established by using BP neural network. The above research has made some achievements in the identification of abnormal user behavior, but the accuracy, robustness, reliability and efficiency of the algorithm need to be improved and improved.

In this paper, K-means clustering method is used to analyze the characteristics and types of each kind of power load. For loads with regular load curves, such as single peak and double peak loads, LES of random matrix theory are used as indicators to monitor power consumption behaviour. When the load fluctuation trend is relatively random, XGBoost regression model, a typical machine learning integrated algorithm, is used to predict the load, and the load curve significantly deviating from the predicted value is regarded as the suspected abnormal power consumption behaviour.

2 Theoretical basis

2.1. K-means clustering algorithm

K-means clustering algorithm is a widely used clustering algorithm, where K is the parameter that needs to be specified. Sum of the squared errors (SSE) or Silhouette Coefficient can be used to determine the value of K. The centroid of k clusters in the K-means algorithm can be obtained in a random way, but these points need to be within the data range.

In the algorithm, the distance from each point to the centroid is calculated, the cluster corresponding to the centroid with the smallest distance is selected as the partition of the data points, and then the centroid of the cluster is updated based on the allocation process. The above process is repeated until the center of mass of each cluster no longer changes. The process of K-means algorithm is shown in Fig. 1.

* Corresponding author: zc2640@sjtu.edu.cn

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

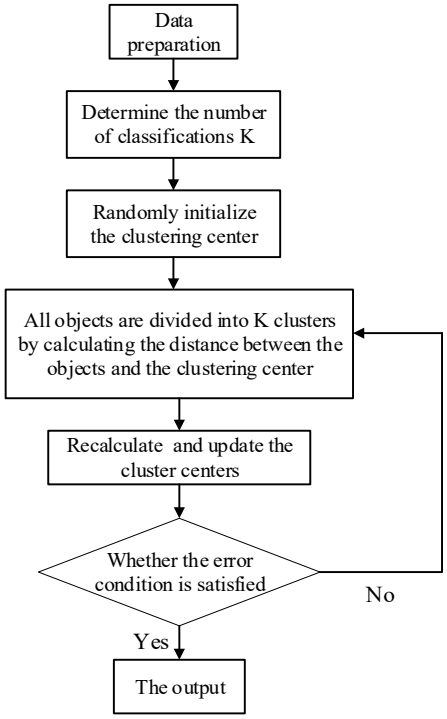


Fig. 1. Training process of K-means clustering algorithm

2.2. Linear eigenvalue statistics (LES)

Random matrix theory (RMT) takes matrix as a unit, and can process the data of independent and identically distributed (IID). According to the stochastic matrix theory, when there is only white noise, small perturbation and measurement error in the system, the data of the system will show a kind of statistical randomness. However, when there is a signal source (event) in the system, the operating mechanism and internal mechanism of the system will be changed under its action, and its statistical randomness will be broken.

For the matrix with N rows and T columns, $\hat{X} \in \mathbb{C}^{N\times T}$ .The transition matrix, $\tilde{X} \in \mathbb{C}^{N\times T}$ can be obtained by the basic transformation formula (1).

$$\tilde{x}_{i,j} = \left(\hat{x}_{i,j} - \mu\left(\hat{x}_i\right)\right) \times \left(\sigma\left(\tilde{x}_i\right) / \sigma\left(\hat{x}_i\right)\right) + \mu\left(\tilde{x}_i\right) \quad (1)$$

where $\mu\left(\tilde{x}_i\right) = 0$, $\sigma^2\left(\tilde{x}_i\right) = 1, i = 1, \cdots, \mathrm{N}$

The singular value equivalent matrix of the transition matrix, $X_u = \mathbb{C}^{N\times N}$ can be obtained by the basic transformation formula (2).

$$X_u = \sqrt{\tilde{X}\tilde{X}^H}U \quad (2)$$

where U is Haar unitary matrix, $X_u X_u{}^H = \tilde{X}\tilde{X}^H$, the superscript H stands for conjugate transpose.

$\tilde{X}_u$ can be obtained by unitizing X_u according to the common formula (3), and $\sigma^2(\tilde{x}_u) = 1/N$.

$$\tilde{x}_{ui} = x_{ui} \Big/ \left[\sqrt{N}\sigma(x_{ui})\right], i = 1, \cdots, N \quad (3)$$

The definition of LES and its statistical characteristics are shown in formula (4):

$$\mathrm{N}_N[\varphi] = \sum_{i=1}^{N} \varphi(\lambda_i) \quad (4)$$

where φ is the continuous test function, λ is the eigenvalue of $\tilde{X}_u$, N is the number of eigenvalues λ .

LES is a statistical feature of the system in a high-dimensional perspective. For a specific system, the threshold value of LES can be given by combining historical data and expert experience. When the LES is not within the range of threshold value, the system will be unstable with a high probability[5]. By establishing a dependency on real-time data random matrix model to create its statistical indicators (such as LES) based on eigenvalues and the theoretical value of this index can be predicted by statistical theories such as The Central Limit Theorem. This results in a real-time data-driven system cognition scheme, whose advantages include data-driven, eigenvalues only, and theoretical support.

2.3. eXtreme gradient boosting (XGBoost)

XGBoost is an optimized algorithm based on Gradient Boosting Decision Tree algorithm (GBDT). GBDT's loss function optimization is a first-order derivative, while XGBoost's loss function optimization is a second-order Taylor expansion, using the first-order derivative plus the second-order derivative [6].

The algorithm not only has the advantages of traditional Boosting algorithm with high accuracy, but also can efficiently process sparse data and flexibly realize distributed parallel computing. Therefore, the XGBoost algorithm is suitable for large data sets. In addition, XGBoost adds a regularization term to the loss function to control the complexity of the model and reduce the variance of the model, making the learned model simpler and preventing the risk of over-fitting. XGBoost improves the accuracy of the algorithm while ensuring a certain computation speed [7].

3 Modelling approach

Modeling method consists of the following steps: 1) Data pre-processing; 2) Clustering; 3) Construction of abnormal behaviour detection model. The modeling flow chart is shown in Fig. 2.

3.1. Data pre-processing

The original sample data contains a large number of vacant data and outliers, which is not conducive to subsequent data analysis. Therefore, batch cleaning of invalid samples should be carried out first. Then the moving window method is used to smooth the data. The magnitudes of power load data are different, so the original data should be standardized before clustering to effectively reduce the impact of data amplitude gap.

3.2. Clustering

K-means algorithm can be used to divide the power users into various types according to the behavior characteristics of electricity consumption, and the user groups with similar behavior characteristics of electricity

consumption can be divided into similar users, which is conducive to the classification and discussion of various user behaviors. In this paper, the typical workday power load data after data preprocessing are used as the feature, the Pearson similarity is used as the distance measure function and SSE is used to determine the value of K.

3.3. Construction of abnormal behaviour detection model.

The power load data was divided into several various types according to the results of clustering algorithm. If the matrix is formed by the power load data obeying a particular distribution, the LES will also show a certain distribution trend. So in this paper, the LES of the matrix is not within the threshold, the system will be unstable with a certain probability. If the power load data does not obey a particular distribution but has a temporal relationship, the XGBoost regression algorithm is used to predict power load data, then set the threshold according to the actual situation. If the actual load data deviates significantly from the predicted value and the difference between the actual and predicted values exceeds the threshold value, the probability of abnormal power consumption behavior is relatively high.

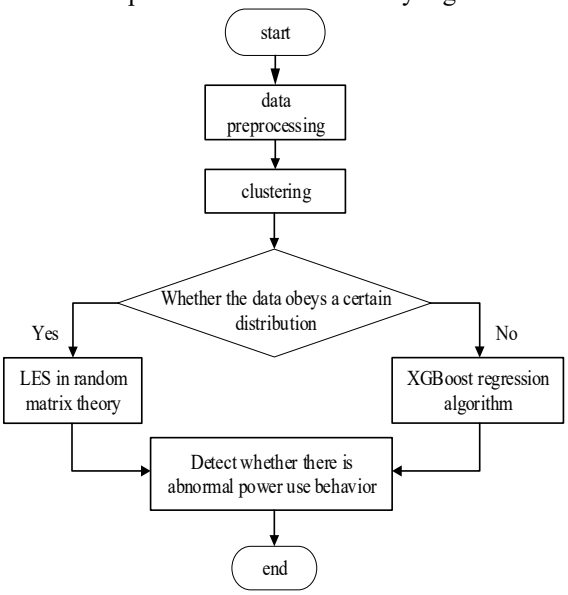
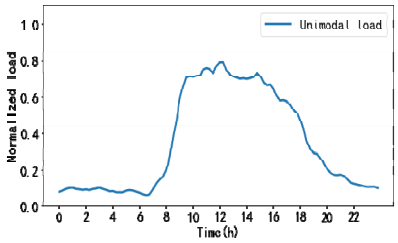


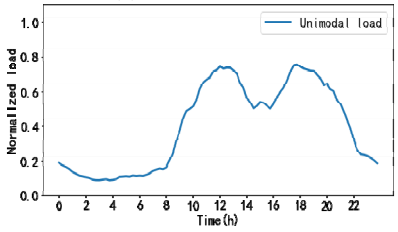
Fig. 2. The modelling flow chart

4 Case studies

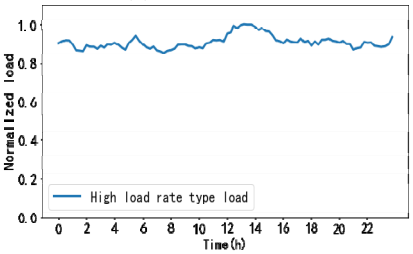
The active power data in this paper came from the measured power load data sampled at 15-minute intervals of an industrial park from 2019 to 2020. The power load data of 98 users are clustered by K-means algorithm. According SSE, 98 power users is divided into 3 types including :1) Unimodal load ;2) Bimodal load ;3) High load rate type load[8].



(a) Unimodal load



(b) Bimodal load



(c) High load rate type load

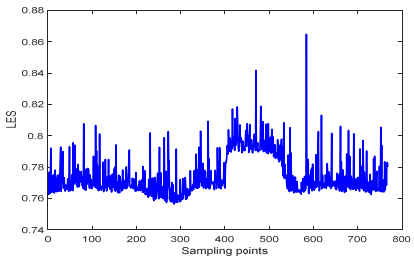
Fig. 3. Three types of power loads

Taking the power load data for 90 consecutive days of a factory which is unimodal peak load, with 96 sampling points per day, namely 8640 data and reshape the data to form a matrix with 10 rows and 864 columns X ,then focus on the matrix Ω_1 in the sliding-window with 10 rows and 96 columns, calculate LES of Ω_1 according to formula (5). Then moving the sliding-window further 1 column to the right and calculate LES of Ω_2 in the sliding-window, Repeat the above steps until the sliding-window moves to the last column of the matrix X.

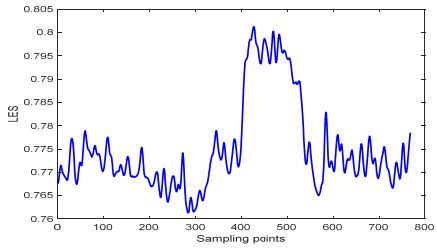
$$N = \sum_{i=1}^{10} \sqrt[8]{|\lambda_i|} \quad (5)$$

where λ is the eigenvalue of X.

Then the LES curve is filtered by the Gaussian filter function in MATLAB. Fig. 4(a) shows the original LES index curve. Fig. 4(b) shows the LES index curve filtered by the Gaussian filter function. One part of curve is U-shaped. This phenomenon indicates abnormal electricity consumption behaviour.



(a) The original LES index curve



(b) The LES index curve filtered by the Gaussian filter function

Fig. 4. The LES index curve

The original power data is shown in the Fig. 5. There is an obvious abnormal electricity consumption behaviour as indicated by the LES index curve.

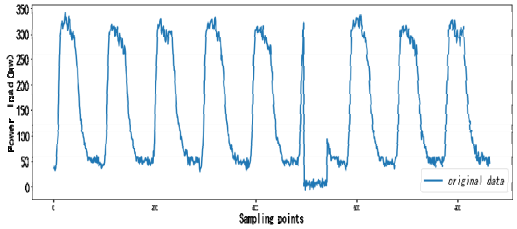


Fig. 5. The original power data

The same method can also be used to detect abnormal electrical consumption behaviour of bimodal load. However, for random fluctuating high load rate type load shown in the Fig. 6, the LES index method is no longer applicable. XGBoost regression model can be used to predict the load. If the load curve significantly deviating from the predicted value, the possibility of abnormal power consumption behavior is relatively high.

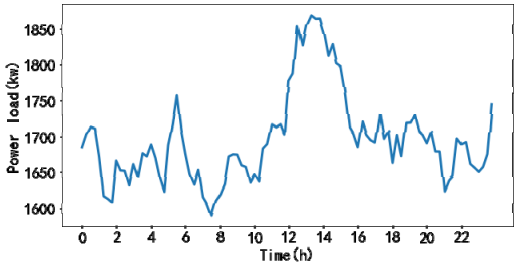


Fig. 6. The random fluctuating high load rate type load

The data of 32 consecutive sampling points were taken as features, and the 33rd sampling point was taken as the value to be predicted. The data from January to November of 2019 is selected as the training set and the February 2019 data is selected as the test set. Then build and train the model based on XGBoost regression algorithm using the training set. The power load prediction results of the model on the test set are shown in the Fig. 7. The model achieves a satisfactory load forecasting effect and provides a basis for judging abnormal power consumption behaviour.

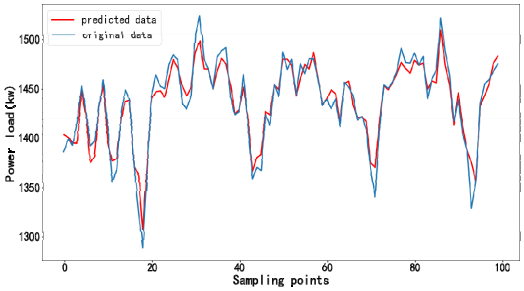


Fig. 7. The power load prediction results on the test set

According to the tolerance of different loads to abnormal electricity consumption behavior, the threshold value of the difference between the actual electricity consumption and the predicted electricity consumption is set to judge whether there is abnormal electricity consumption behavior.

5 Conclusion

This paper is based on the measured big data of power grid. The power loads are divided into load types that obey specific distribution law or with random fluctuation.

The LES index can be used as the tool to detect the abnormal power consumption behavior for that load types obeying specific distribution such as unimodal load or bimodal load. For random fluctuating high load rate type load, the difference between the actual and predicted value of XGBoost regression model can be used as the basis for judging abnormal power consumption behaviour. The method proposed in this paper can detect abnormal behaviour of different types of loads and implement a good discriminant effect.

Fund project

State Grid Shanghai Electric Power Company Science and Technology Project (Research on Measurement Monitoring Statistical Characteristics Analysis and Evaluation Technology Based on Complex Perception Model -520940190078)

References

1. S. ZHENG, Q. LIANG, X. PENG, et al.Research on Recognition of Abnormal Electrical Behavior Based on Fuzzy Clustering [J]. Electrical Measurement & Instrumentation, 2020(**19**):40-44
2. M. Lin, X. Peng, L. Lin, et al. Electricity price execution online audit model based on data mining technology. Guangdong Electric Power. 2016, **29(01)**:108-112.
3. C. Zhuang, B. Zhang, J. Hu, et al. Abnormal power consumption pattern detection of power users based on unsupervised learning. Proceedings of the CSEE. 2016, **36(2)**:379-387.
4. Z. Cao, J. YANG, X. LIU. Study and application of preventing system from stealing power based on BP

neural network [J]. Water Resources and Power, 2011, **29 (9)**: 199-202.

5. X. HE, Q. AI, C. Qiu, et al. Application of Stochastic Matrix Theory in Power System Cognition [J]. Power System Technology, 2017, 04(v.**41**; No.401):151-159.
6. H. Meng, X. Huang, X.Tu.Throughput prediction of steel coil storage based on time series and XGBoost [J]. Application of Computer, 2019(**S02**):24-28.
7. L. REN, L.ZHANG, H. WANG, et al. IXGBOOST Short-term Power Load Prediction Based on Optimal Clustering [J]. Computer and Digital Engineering, 2020, v.**48**; No.366(04):12-18.
8. Y.Ma, J. Chen, L. Hua, et al. Morphological analysis and evaluation of power demand side load [J]. Mechanical and Electrical Information, 2015 (**24**) :18-21.