Analysis and Application of Abnormal Electricity Based on Mean Shift Clustering and XGBoost Verification

Aimin Liu², Feng Sun¹, Wei Li², Xin Wen², Tong Wang¹, Xuke Cheng*¹

¹ Electrical Power Research Institute of Liaoning Electric Power Co., Ltd., Shenyang 110006, China
² Liaoning Electric Power Co., Ltd, Shenyang 110006, China

*Mail Address: cxk360@126.com

Abstract. With the gradual development of consumer behavior analysis, abnormal electricity consumption analysis has become a hot topic in the analysis of mining data. At present, the abnormal power analysis method based on manual verification and judgment is inefficient and has a low hit rate. In this paper, the mean shift clustering algorithm is adopted to cluster the electricity consumption and the fluctuation of electricity consumption respectively and the residents with large electricity consumption and large fluctuation of electricity consumption are selected as suspected abnormal electricity consumption. Then, based on the decision tree model of xgboost, the users suspected of abnormal electricity consumption are filtered twice to realize the automatic study and judgment of electricity consumption behavior and make full use of the massive data resources of the power grid. The value of the source greatly improves the efficiency of verification and helps power enterprises to recover considerable economic losses.

1. Introduction
Abnormal electricity consumption is inevitable among the huge users of daily electricity consumption. For example, the resident users privately use the power load at the metering point of residential electricity consumption for commercial production or operation, which brings about potential safety hazards and economic losses of power enterprises. At present, with the realization of the goal of full coverage, full cost control and full collection of electricity information collection system by power enterprises, the monitoring of low-voltage stations is getting more and more in-depth by utilizing the massive data resources of mining system. However, abnormal power consumption behavior of low-voltage users in the station area cannot be controlled in time, and cannot fully meet the urgent demand of marketing management for intelligent diagnosis of user-side power consumption behavior.

It has become an urgent need for the marketing management of low-voltage power stations to monitor and analyze the users' default power consumption by using the existing daily frozen power consumption data.

The traditional method of checking the breach of contract mainly relies on manual inspection by strengthening the sense of responsibility consciousness of inspectors and installing recorders, etc. [1] However, this method is inefficient and overly relies on the subjective consciousness of personal, and the checking takes randomness, the hit rate is too low; Keyan Cao et al. [2] proposed local outliers based on specific uncertain data model. The basic is proposed, and effectively the calculation is reduced the amount of calculation by estimating the local monitoring algorithm; residential lighting...
users freeze abnormal points in China and Japan, Wang Yong et al. [3] use clustering algorithm to compare and summarize different types of data curves, and classified the data with similar rules into the same category; user power consumption is high-dimensional data, for high-dimensional data. Papadimitriou S et al. [4] proposed an outlier estimation method based on local correlation integral method, which provides a trajectory diagram for each point and summarized abundant information about the data near the point, the determination of clusters and micro-clusters, their diameters and the distance between clusters. Jiang M F et al. [5] proposed a two-stage clustering algorithm for outlier detection. This method improves the traditional K-means algorithm and adopts the heuristic clustering center method. The experimental results show that this method has a good process operation. Yoon K et al. [6] proposed a K-means clustering method Zhou Y et al. [7] clustered high-dimensional data in three stages to find outlier. In the first stage, clustering cluster was initially determined. Local outliers of each cluster are identified and their effects on the centroid are eliminated in the second stage. Therefore, the outlier is finally determined. experimental results show that the algorithm supports clustering with different densities, sizes and non-spheres. Kim S S. et al. [8-9] can roughly distinguish the power users who are different from ordinary users by using K-means clustering method to identify anomalies or single groups, but this method is not accurate enough to locate accurately, and the execution efficiency needs to be improved under large data volume. Tomas Vojir et al. [10] use Mean-Shift algorithm to track target. Most of the above methods adopt K-means algorithm for analysis, but in the face of a large amount of power consumption data, k-means requires iterative clustering to obtain the optimal number of clusters, so the mean shift algorithm is adopted for clustering.

This paper analyzes the advantages and disadvantages of all parties, and uses the user's daily frozen electricity data for analysis. Due to the uncertainty of the number of clusters, the mean shift clustering method was used instead of K-means clustering method to cluster the power consumption and fluctuation amplitude respectively, and the residents with large power consumption and fluctuation amplitude were identified as suspicious abnormal users. In order to further narrow the suspected range and improve the accuracy of the results, the decision tree model based on xgboost is adopted to build the model with the data of known abnormal users. The results of the above clustering intersection were filtered twice to generate the final list of abnormal users, which was used by business personnel for manual checking. With the method proposed in this paper, the accuracy is over 80%, which is obviously improved compared with other methods.

2. Model Building and Application

2.1. Main steps of model building

The construction steps of the abnormal power consumption model are described as follows:

Step 1: Collect the daily frozen power consumption information of users in the station area, and convert the date to column label by row and column transposition.

Step 2: Cluster analysis of mean shift based on electricity consumption information and normalized electricity fluctuation;

Step 3: Identify and intersect the difference of clustering results in step 2 to form the initial suspected abnormal electricity list.

Step 4: 80% of the confirmed abnormal power list as samples for training and learning, and form a decision tree model based on electricity data. The model is used to verify the remaining 20% abnormal power consumption data and continuously optimize the adjusted decision tree model.

Step 5: Use the model of step 4 to screen the suspected abnormal electricity list of step 3 twice, and get the final abnormal electricity list.
2.2. Clustering of Power Consumption and Electricity Volatility

Mean Shift algorithm is a non-parametric method based on density gradient rise. It is often used for object tracking, data clustering, classification and other scenes of image recognition.

The core idea is: firstly, select a center point at random, then calculate the average distance vector pointing to the center point within a certain range, calculate the average value to get an offset mean, and then move the center point to the offset average position. Through this repeated motion, the center point can gradually approach the optimal position. This idea is similar to the gradient descent method. By moving continuously in the direction of gradient descent, the local or global optimal solution on gradient can be obtained.

With the help of the mean shift clustering algorithm, residential lighting users with large power consumption and large fluctuation. Their electricity consumption characteristics are very similar to those of non-residential users. It is very likely to receive non-residential load at home, production and operation, abnormal electricity consumption.

The specific clustering steps are as follows, firstly, data size and fluctuation amplitude are used to cluster the data. The fluctuation amplitude curve needs to normalized to eliminate the influence of the size difference on the fluctuation, and then clustering is carried out. After clustering, the residents in the curve with the highest level of power consumption are set as a set; the residents in the curve with the largest fluctuation of power consumption are taken as a group; The suspected default users is set as the set is the intersection of sets and sets.

The formal description of the whole algorithm is shown in algorithm 1.

Input: daily frozen electricity by users of low-voltage stations;
Output: A collection of suspects with large power consumption and large fluctuation range;
- Input the daily frozen power of users in low-voltage station area;
- Clustering by means of mean shift method, the central curve of electricity consumption is formed.
- Normalize the power consumption data for each user, and then use the mean shift method to cluster to form the power fluctuation center curve.
- The resident user in the curve with the highest level of electricity consumption is set as C1.
- The resident user in the curve with the greatest fluctuation of electricity withdrawal is set as C2.
- The final output suspect list is set as a set, and the set is the intersection of set C1 and C2.

On this basis, a list of residential lighting users with large power consumption and large fluctuation is obtained. These users have high suspicion of default power consumption in the sample.
2.3. Decision Tree Model Based on XGBoost

XGBoost is one of the boosting algorithms. The idea of Boosting algorithm is to integrate multiple weak classifiers into one a strong classifier. Since XGBoost is a lifting tree model, it integrates many tree models to form a strong classifier. The tree model used is the CART regression tree model.

In this paper, we select 50 electricity records of the users who have been identified as abnormal power consumption. Among the users who have been identified as non-abnormal power consumption, 50 electricity records, totaling 100 records, are selected as input samples for model training. The output is 1 and 0. 1 indicates that abnormal power consumption is determined and 0 is non-abnormal power consumption.

Softmax objective function was used for training. Since the result of classification is abnormal and normal, the number of num_class classes is set to 2. Python is basically implemented as follows:

```python
xgb_model = xgb.XGBClassifier(objective="multi:softmax", num_class=2)
xgb_model.fit(train, label.values.ravel())
```

Among them, the first parameter of fit method is the sample data to be trained, and the second parameter is the output of the corresponding category or value, which is designated as true and false. true represents the abnormal user, while false is the non-abnormal user. Generate model pkl file.

Next, the initial default user's power consumption generated in the previous step is taken as input, and the xgboost model file is invoked, and the user whose result is abnormal power consumption is taken as the final output.

3. Model Application Verification

3.1. Data preparation

In this paper, daily frozen power data of about 96 stations in Dalian City of Liaoning Electric Power Company in 2016 are extracted. In the database, data is stored in a row-by-row format, and a user records one row per day. When building the xgboost model, Liaoning Electric Power Company provided 100 users in the province who were identified as abnormal users and 100 users who were identified as non-abnormal users for a long time. The power consumption data format is as follows:

| NO | TQBH | YHHH | YHMC | DATE       | … | POWER |
|----|------|------|------|------------|----|-------|
| 1  | 06900XXXXX | XXX | XX  | 2017/01/01 | … | 4.9   |
| 2  | 06900XXXXX | XXX | XX  | 2017/01/02 | … | 5.1   |
| …  | …   | …   | …   | …          | … | …     |
| 365| 06900XXXXX | XXX | XX  | 2017/12/31 | … | 4.8   |

3.2. Clustering

This paper analyzes the data of a station in Dalian as an example. The station contains 723 users, including 607 resident users, 116 non-resident users, and the proportion of non-resident users is 16%.

Firstly, power consumption data is clustered based on mean shift, which is one year data, so it is 365 dimensions. The clustering results are shown in Figure 2.
As shown in Figure 2 and Figure 3, clustering center point curves are divided into three categories. There are 101 users in the first category, including 65 resident users and 36 non-resident users; There are 640 users in the second category, including 549 resident users and 91 non-resident users, who are excluded due to low electricity consumption; and 11 users in the third category, all of which are non-resident users, help to exclude them. Therefore, 101 users in the first category are regarded as suspected objects in the dimension of electricity consumption.

The power consumption data are normalized and clustered to obtain the power consumption fluctuation clustering curve as shown in the figure.

Figure 2. Clustering Central Point Curve of Electricity Consumption

Figure 3. Clustering Quantity of Electricity Consumption Categories

As shown in Figure 2 and Figure 3, clustering center point curves are divided into three categories. There are 101 users in the first category, including 65 resident users and 36 non-resident users; There are 640 users in the second category, including 549 resident users and 91 non-resident users, who are excluded due to low electricity consumption; and 11 users in the third category, all of which are non-resident users, help to exclude them. Therefore, 101 users in the first category are regarded as suspected objects in the dimension of electricity consumption.

The power consumption data are normalized and clustered to obtain the power consumption fluctuation clustering curve as shown in the figure.

Figure 4. Clustering Central Point Curve of Electricity Volatility
As shown in Fig. 4 and Fig. 5, the cluster centers of power fluctuations is divided into two categories: 464 users in the first category and 288 users in the second category. The fluctuation range of the second category is larger than that of the first. The second category is selected as the suspected object of power fluctuation dimension.

Combined with the results of two clustering, 39 users with suspected abnormal power consumption were obtained.

3.3. Decision Tree Analysis

The model is used to screen and filter the 123 users mentioned above. Finally, 21 checklists of abnormal users are obtained. According to the 21 checklists, on-site manual checks are carried out. The results show that 16 of the 21 users have commercial power load, and the abnormal electricity is authentic. The checking accuracy is 81.16%.

4. Results and Conclusions

In this paper, clustering and decision tree algorithm are used to diagnose abnormal power consumption behavior of a large number of low-voltage users, aiming at daily frozen electricity data. The annual economic benefit of excavation is more than 30 million yuan, and the average annual economic benefit of default users reaches 11.5 million yuan. Compared with the traditional manual spot checking method, the analysis model in this paper greatly improves the monitoring area and efficiency of default power consumption, fully reduces the of power consumption management and control, and makes up for the business defects of power enterprises which are difficult to manage and control the power consumption behavior of users in low-voltage stations with large quantities and high efficiency.

References
[1] Wang Tian-an. Management Measures of Electricity Stealing and Default in Electricity Inspection[J]. Technological Economy and Management, 2016, (09): 140.
[2] Cao Ke-yan, Luan Fang-jun, Sun Huan-liang, Ding Guo-hui. Density-Based Local Outlier Detection on Uncertain Data[J/OL]. Chinsese Journal of Computers, 2016,:1-15(2016-5)
[3] Wang Yong, Tang Jing, Rao Qin-fei, Yuan Chao-yan. High efficient K-means algorithm for determining optimal number of clusters[J]. Journal of Computer Applications , 2014, (05): 1331-1335.
[4] Papadimitriou S, Kitagawa H, Gibbons P B, et al. LOCI: fast outlier detection using the local correlation integral[C]. International Conference on Data Engineering, 2003. Proceedings. IEEE, 2003:315-326.
[5] Jiang M F, Tseng S, Su C M. Two-phase clustering process for outliers detection[J]. Pattern Recognition Letters, 2001, 22(6):691-700.
[6] Yoon K, Kwon O S, Bae D H. An Approach to Outlier Detection of Software Measurement Data using the K-means Clustering Method[C]. International Symposium on Empirical Software Engineering and Measurement. IEEE Computer Society, 2007:443-445.

[7] Zhou Y, Yu H, Cai X. A Novel k-Means Algorithm for Clustering and Outlier Detection[C]. International Conference on Future Information Technology and Management Engineering. IEEE, 2010:476-480.

[8] Kim S S. Variable Selection and Outlier Detection for Automated K-means Clustering[J]. Communications for Statistical Applications & Methods, 2015, 22(1):55-67.

[9] Songma S, Chimphlee W, Maichalernnukul K, et al. Classification via k-means clustering and distance-based outlier detection[C]. International Conference on ICT and Knowledge Engineering. IEEE, 2013:125-128.

[10] Tomas Vojir, Jana Noskova, and Jiri Matas. Robust Scale-adaptive Mean-Shift for Tracking[C] SCIA 2013.