Abnormal behavior analysis of electricity consumption based on improved random forest with grey relation projection

WANG Kun\textsuperscript{1,2}, GAO Jinggeng\textsuperscript{1}

\textsuperscript{1}State Grid Gansu Electric Power Company Marketing Service Center, Gansu Lanzhou, 730050

E-Mail: wangkun@gs.sgcc.com.cn

Abstracts: With the construction of ubiquitous power Internet of things and the deepening application of big data technology, it is possible to mine the abnormal behavior of electricity consumption from the massive power consumption information. In view of the current application of anti stealing power in power system, an improved random forest algorithm based on grey relation projection is proposed to analyze the abnormal behavior of power consumption. The improved stochastic Sen algorithm is used to build the diagnosis model of abnormal electricity consumption of users, and analyze the power data of the power consumption information acquisition system, so as to improve the accuracy of the intelligent diagnosis algorithm of anti stealing electricity. The effectiveness of the proposed algorithm can be obtained by verification test.

1. Introduction

With the deepening of China's electric power reform and the vigorous development of electricity spot trading market, power generation, power supply enterprises and power users' requirements for power supply reliability of power system continue to improve, and the importance of accuracy and fairness of electric energy measurement is increasingly strengthened\cite{1,2}. The in-depth study on the diagnosis analysis and model building technology of power consumption data on the user side is conducive to strengthening the demand side management of power grid, reducing the non-technical loss of power enterprises, and building a safe and strong smart grid\cite{3,4}. The traditional detection technology of abnormal behavior of users can be roughly divided into three categories: probability and statistics method, calculation distance method and machine science method\cite{5-7}. Theoretical methods such as support vector machine (SVM), neural network (neural network), decision tree and other artificial intelligence methods have achieved good results in the field of abnormal electricity use monitoring. With the deepening of power system intelligence, various data information such as electric quantity and non-electric quantity collected by electronic intelligent watt hour meter and sensor show a power exponential growth trend with high complexity and high redundancy. Due to the large number of power users in China, the complexity of power grid user side data increases, which gradually constitutes the user side big data. In the face of the complex characteristics of power consumption big data information, the traditional power consumption behavior mode detection technology has been unable to meet its analysis and processing requirements, so the analysis and processing technology problems of user electricity consumption behavior big data need to be solved urgently.

In recent years, experts and scholars at home and abroad have also conducted in-depth research on the detection technology of abnormal behavior of electricity consumption. In reference [8], a detection method of abnormal behavior of power consumption side based on fuzzy set theory and cluster
analysis is proposed. However, the above scheme has single judgment mode, less data feature association and insufficient data mining of abnormal electricity consumption behavior.

In order to make full use of the huge amount of electricity consumption behavior data accumulated by power enterprises and build a more accurate identification model of abnormal electricity consumption behavior, this paper proposes an analysis method of abnormal electricity behavior based on Grey projection improved random forest algorithm. The improved random forest algorithm is used to establish the abnormal model of electricity consumption behavior, and the sample set selected by grey projection is used to train the model, so as to get more accurate evaluation of abnormal electricity use behavior.

2. Selecting sample set by weighted grey relational projection method
The weighted grey relational projection method is composed of grey system theory and vector projection principle. This method overcomes the disadvantage of only using grey correlation coefficient to evaluate the correlation degree of samples. The concept of weighted sum projection is introduced. Firstly, the key influencing factors are highlighted by using appropriate weighting method, and then the projection value of historical samples on the estimated samples is used to comprehensively evaluate the correlation degree between historical samples and samples to be estimated, and a data set similar to the estimated data is obtained.

(1) Firstly, m factors closely related to electricity consumption behavior are selected, then the eigenvector of the ith sample can be expressed as follows:

\[ Y_i = [y_{i1}, y_{i2}, \ldots, y_{im}] \quad i = 1, 2, \ldots, n \]  

Where n is the total number of historical samples and the \( y_{im} \) influencing factor value of the ith sample.

(2) According to the feature vector, the grey correlation judgment matrix is established, and the correlation coefficient between the sub sequence and the parent sequence is calculated:

\[ F = \begin{bmatrix} F_{01} & \cdots & F_{0m} \\ \vdots & \ddots & \vdots \\ F_{n1} & \cdots & F_{nm} \end{bmatrix} \]  

Where \( F_{nm} \) is the grey correlation degree value of the m factor of the nth sample.

(3) The weight of each influencing factor is determined by entropy weight method:

\[ W = [w_1, w_2, \ldots, w_m] \]  

Where \( w_m \) is the weight value of the m influencing factor.

(4) The weighted grey relational decision matrix is obtained by weighting the above weight vectors:

\[ F' = FW^T = \begin{bmatrix} w_1 & \cdots & w_m \\ \vdots & \ddots & \vdots \\ w_1F_{n1} & \cdots & w_mF_{nm} \end{bmatrix} \]  

(5) If each row in the correlation decision matrix is regarded as a row vector, the angle of each row vector is the gray projection angle of the sample. Thus, the gray correlation projection value is:

\[ D_i = \frac{\sum_{j=1}^{m} w_j F_{ij} w_j}{\sqrt{\sum_{j=1}^{m} (w_j F_{ij})^2} \sqrt{\sum_{j=1}^{m} w_j^2}} \]
Where $D_i$ is the projection value of the ith sample vector on the sample vector to be estimated.

(6) According to the gray projection value of each historical data vector, the projection value threshold is set, and the sample with larger projection value is selected as the sample set.

3. random forest algorithm

Random forest algorithm is a supervised ensemble learning algorithm. Its core idea is to combine multiple classification and regression trees (CART) with weak performance into a forest through certain rules. The result is obtained by voting all decision trees in the forest.

3.1 Classification and Regression tree

Classification and regression tree is a binary recursive segmentation technology, which divides the current sample set into two subsets on each node except leaf node. The Gini index is the measure of attribute selection in cart algorithm. Assuming that data set D contains M categories, the Gini index can be obtained as follows:

$$G_D = 1 - \sum_{j=1}^{M} p_j^2$$

Where $p_j$ is the frequency of J-type elements.

The Gini index needs to consider the binary partition of each attribute. Assuming that the binary partition of attribute $a$ divides data set $D$ into $D_1$ and $D_2$, then the Gini index of sample set $D$ divided by attribute $a$ in sub node is as follows:

$$G_{D_{1},a} = \frac{|D_1|}{D} G_{D_{1}}(D_{1}) + \frac{|D_2|}{D} G_{D_{2}}(D_{2})$$

For each attribute, considering every possible binary partition, the subset of the minimum Gini index produced by the attribute is selected as its split subset. Therefore, the smaller the Gini index $G_D$ and $a$ on attribute $a$, the better the partition effect is. Under this rule, the decision tree is divided from top to bottom until the growth of the whole decision tree is completed.

3.2 Bagging method and random attribute subspace sampling method

Because of the limited prediction accuracy of cart decision tree, it can not meet the requirements of power consumption anomaly analysis and evaluation. Therefore, bagging algorithm is used to improve the prediction accuracy. In this method, bootstrap repeatable sampling is used to extract sub training sets of equal scale from the original training set for each cart tree. The results show that the method can improve the research ability of the classifier. At the same time, when each node of cart tree is split, several attributes are randomly extracted to form attribute subspace to split. Bagging method enhances the performance of single decision tree in forest, while attribute subspace sampling method reduces the correlation between each tree.

The upper bound of random forest generalization error is as follows:

$$E^* \leq \bar{\rho}(1 - s^2)$$

Where $s$ and $\bar{\rho}$ are the average correlation coefficient and average strength of the tree. Therefore, with the decrease of tree correlation and the increase of single tree strength, the upper bound of generalization error of random forest will be reduced, and its generalization error will be effectively controlled. Therefore, the main ways to improve the accuracy of random forest prediction are as follows: reducing the correlation of trees and improving the performance of single classifier (i.e. single decision tree).

4. Weighted grey projection improved stochastic forest algorithm process

To sum up, the calculation steps of the improved random forest algorithm based on weighted grey
relational projection are as follows:

1. Firstly, the user power consumption information set is initialized, and the weighted grey
   relational projection method is used to form the training sample set with high similarity;
2. The training set is resampled by bootstrap to generate the sub training set;
3. According to the algorithm in Section 2.2, cart decision tree is generated, and the number of
   randomly selected features is determined by the scale of input information set;
4. According to the random forest model, the average output value of each tree is calculated, and
   the judgment value of abnormal behavior is obtained by comparing with the penalty threshold of
   abnormal accumulation.

5. Simulation Analysis
In this paper, the validity of the model is verified by using the standard data set and the data of one
year power consumption information acquisition system in a certain area of Gansu Province.
Firstly, the classification characteristics of the proposed algorithm are verified. The experiment
uses BP neural network as the control group, in which the number of hidden layer neurons of BP
neural network is 24; the improved random forest algorithm uses penalty factor 0.06, the number of
decision trees is 21, and the resampling ratio is 0.17. The simulation results are shown in Table 1.

| Algorithm                | Comparison | Simulation |
|--------------------------|------------|------------|
| BP neural network        | Accuracy   | 95.42%     |
|                          | RMSE       | 0.2475     |
|                          | Training time | 1.945s     |
| Improved random forest   | Accuracy   | 98.37%     |
|                          | RMSE       | 0.4215     |
|                          | Training time | 2.5364s    |

As shown in Table 1, the accuracy of the simulation results obtained by the improved random forest
algorithm is higher than that obtained by the BP neural network. At the same time, the improved
random forest algorithm suppresses the over fitting phenomenon in the traditional random forest
algorithm, so the RMSE value is not more than that of the BP neural network algorithm. It can be
concluded that the classification accuracy of the proposed algorithm is higher than that of the BP
neural network algorithm.

According to the Hadoop distributed experimental platform, the abnormal electricity consumption
behavior of the above two algorithms is analyzed and tested, excluding the factors such as work ticket
delay corresponding to the power consumption information acquisition system and marketing system.
The power consumption information without abnormal maintenance records in the system is used as
the training sample, and the k-means algorithm is used as the control algorithm. The simulation results
are shown in Table 2, in which the detection rate is the test The ratio between the number of abnormal
samples detected and the total number of abnormal samples in the trial set. The false detection rate is
the ratio between the number of normal samples misjudged as abnormal and the total number of
samples detected.

| Algorithm        | Comparison   | Simulation |
|------------------|--------------|------------|
| K-means          | Detection rate | 72.61%     |
|                  | Noise factor  | 19.24%     |
|                  | Training time | 51.28s     |
| Improved         | Detection rate | 92.53%     |
| random forest | Noise factor | 5.42% |
|---------------|-------------|-------|
|               | Training time | 94.33s |

It can be seen from the experimental results shown in Table 2 that the simulation results obtained by the improved random forest algorithm are significantly better than the k-means algorithm in terms of detection rate and false detection rate. Therefore, the calculation results of the proposed algorithm in power consumption anomaly analysis meet the actual requirements.

6. Conclusion
This paper presents an analysis method of abnormal electricity consumption behavior based on Grey projection improved random forest algorithm. The improved random forest algorithm is used to establish the abnormal model of electricity consumption behavior, and the sample set selected by grey projection is used to train the model, so as to obtain more accurate analysis results of abnormal electricity use behavior.

Acknowledgement
This work is support by Science and Technology Project of State Grid Gansu Electric Power Company (No. 52273118000Y).

Reference
[1] León C, Biscarri F, Monedero I, et al. Variability and trend-based generalized rule induction model to NTL detection in power companies[J]. IEEE Transactions on Power Systems, 2011, 26(4): 1798-1807.
[2] Lin Nvgui, Hong Lanxiu, Huang Daoshan, et al. Identification of abnormal electricity consumption behavior based on improved deep self coding network[J]. China Electric Power, 2020, 53(06): 18-26.
[3] Zhuang Chijie, Zhang Bin, Hu Jun, et al. Detection of abnormal power consumption mode of power users based on unsupervised learning[J]. Proceedings of the CSEE, 2016, 36(2): 379-387.
[4] Song Yaqi, Zhou Guoliang, Zhu Yongli, et al. Status and challenges of big data processing technology in Smart Grid [J]. Power System Technology, 2013, 37(4): 927-935.
[5] Zhang Tiefeng, Gu Mingdi. Power user load mode extraction technology and its application [J]. Power System Technology, 2016, 40(3): 804-811.
[6] Monedero I, Biscarri F, León C, et al. Detection of frauds and other non-technical losses in a power utility using Pearson coefficient, Bayesian networks and decision trees [J]. International Journal of Electrical Power and Energy Systems, 2012, 34(1): 90-98.
[7] Cheng Chao, Zhang Hanjing, Jing Zhimin, et al. Research on anti stealing electricity based on outlier algorithm and power consumption information acquisition system [J]. Power System Protection and Control, 2015, 43(17): 69-74.
[8] dos Angelos E W S, Saavedra O R, Cortes O A, et al. Detection and identification of abnormalities in customer consumptions in power distribution systems[J]. IEEE Transactions on Power Delivery, 2011, 26(4): 2436-2442.