Online cleaning method of power grid energy anomaly data based on improved random forest

Ke Chen1*, Hongkai Wang1, Zhangchi Ying1, Chengxin Zhang1, Jiaqi Wang1

1State Grid Zhejiang Electric Power Corporation Information & Telecommunication Branch, Hangzhou, Zhejiang, 310016, China

*ricemice1227@163.com

Abstract—Aiming at the problem of high root mean square error of traditional power grid energy anomaly data online cleaning, a power grid energy anomaly data online cleaning method based on improved random forest is designed. Firstly, an outlier data recognition model of isolated forest is designed to identify outliers in the data. Secondly, an improved random forest regression model is established to improve the adaptability of random forest to mixed abnormal data, and the data trend is fitted and predicted. Finally, the improved random forest data cleaning method is used to compensate the missing data after removing the mixed abnormal data, so as to clean the abnormal energy data of the power grid. The experimental results show that when the amount of power grid energy anomaly data increases, the cleaning root mean square error of the experimental group is significantly lower than that of the control group. The method in this paper solves the problem of high root-mean-square error in the online cleaning of abnormal data of traditional grid energy.

1. Introduction

The power grid contains huge real-time data, which is closely related to the safe operation of the power system and plays a decisive role in the safety and reliability of the power grid. The power grid needs to generate and consume a lot of electric energy in the process of working. Under the condition of normal work, it will still produce bad data, affect the decision-making of the system and threaten the safety of the whole power grid. Data cleaning is not only applied in power grid, but also in the fields of data mining and total data quality management. The definitions of data cleaning are different in different fields, and there is no unified definition of data cleaning. Therefore, using professional technology to detect and correct bad data will help to ensure the normal operation of power grid and the normal supply of power energy. So far, scholars at home and abroad have done a lot of theoretical and practical research on the detection and correction of power grid energy big data, and summarized two widely used cleaning methods. One is the identification and cleaning method of abnormal data of power grid operation based on Spark framework. Abnormal data are extracted from big data through data mining algorithms such as clustering or neural network, and cleaned under Spark framework. The other is the missing value cleaning method of power grid data based on improved low rank matrix completion. This method mainly sets the threshold according to a certain confidence level, carries out hypothesis test combined with the relevant knowledge of probability theory, and then completes the missing value of data by improving the low rank matrix to realize anomaly cleaning. In the actual operation of the power grid, it is very difficult to ensure the accuracy and completeness of user electricity data collection. Affected by sensor failures, transmission line failures, natural weather and other complex factors, users’ electricity consumption data will have varying degrees of data omissions.
and errors. This article refers to these user electricity consumption data as abnormal data. In the above two methods, the data recognition is abnormal due to too many repeated searches, or the normal data is judged as bad data, resulting in a high root mean square error in practical applications.

The data set does not need standardization, and the training speed is fast. It can be applied to large-scale data sets. Default values can be processed (as a separate class) without additional processing. Due to the out of bag data (OOB), the unbiased estimation of the real error can be obtained in the process of model generation without losing the amount of training data. In the training process, the interaction between features can be detected, and the importance of features can be obtained, which has a certain reference significance. Because each tree can be generated independently and simultaneously, it is easy to make a parallelization method. Because of its simple implementation, high precision and strong anti overfitting ability, it is suitable to be used as a benchmark model in the face of nonlinear data. Based on this, this paper designs an online cleaning method of power grid energy anomaly data based on improved random forest to solve the problems existing in the previous cleaning methods.

2. Improved random forest

In order to adapt to the compensation of mixed missing data, the random forest algorithm is improved to obtain an improved random forest algorithm with strong adaptability [1-2]. The X excluding abnormal data is linearly interpolated, and then the filling matrix containing target compensation variables is obtained through matrix transformation. Then, the random forest regression model is used to predict the filling matrix, and the integration idea is used to output $t_b$ times\(^3\)\(^4\). The specific methods for improving random forest are as follows:

In step 1, the missing value in X is interpolated by linear interpolation method to obtain matrix R, which is expressed as:

$$R = [x_1, x_2, x_3, ..., x_n]$$ (1)

In formula (1): the sample size of R and X is the same, which is $m \times n$ dimensional matrix, $n$ is the number of variables, and $m$ is the number of data samples contained in a single variable.

In step 2, take column i in X as the target filling column, and the remaining n-1 column in R except column i as the relevant variable column to form the filling matrix R$^i$, \(I = 1, 2, 3,..., n\), expressed as:

$$C(u) = 2(In(u-1) + \xi) - \frac{2(u-1)}{u}$$ (2)

In step 3, a training matrix $D_\nu$ is randomly selected from D to complete the establishment of random forest regression model.

In step 4, take the relevant variable column in R$^i$ as the input and the target filling column as the output to predict the missing value of column i to obtain the $t$-th prediction value $y(t)$, $t = 1, 2, 3,..., t_b$.

Step 5 if $t < t_b$, then $t = t + 1$, return to step 4; otherwise, take the average value of $t_b$ predictions as the final compensation value $y_\nu$, and replace the missing value in $x_i$ with $y_\nu$ to complete the update of matrix X.

$$y_\nu = \frac{1}{t_b} \sum_{i=1}^{t_b} y(t)$$ (3)

Step 6 if $i < n$, $i = i + 1$, repeat the above steps; otherwise, stop prediction and complete data compensation for X\(^5\)\(^6\). The above is to improve the overall step flow of random forest algorithm.

3. Online cleaning method of power grid energy anomaly data based on improved random forest

3.1. Classification of grid energy anomaly data

According to the analysis of the original data of power grid energy, the abnormal data of power grid energy are mainly divided into the following three categories: outlier data, duplicate data and missing data. Outliers are data that deviate significantly from the rest of the values\(^7\). Duplicate data is the data
with the same value at multiple times. According to the dynamic characteristics of power grid operation process, this kind of data is determined as abnormal data. Missing data refers to data not collected at a certain time or data vacancy\(^8\). However, there are many energy data variables and parameters in the power grid, and the abnormal data do not appear in a single type, but often appear in a variety of abnormal characteristics at the same time. At present, the research on online data cleaning of power grid energy anomalies is limited to cleaning methods for single type of abnormal data, and there is no mixed abnormal data cleaning method for different types. When the traditional probability distribution, clustering algorithm and intelligent algorithm are used to clean different types of mixed abnormal data, the effect is still difficult to meet the requirements of data quality during power grid operation. Therefore, this paper proposes a data cleaning method based on improved random forest to clean different types of mixed abnormal data.

### 3.2. Identification of grid energy anomaly data based on improved random forest

In order to quickly and accurately identify abnormal data, an abnormal data identification model based on isolated forest is constructed. IF is a decision tree ensemble learning method. IF algorithm first divides the known and continuous time data set randomly, and then realizes fast and accurate identification by using the difference between abnormal data and normal data. The IF model consists of a isolation tree, and the algorithm steps are as follows:

1. **Step 1** randomly select a sub sample matrix \( X_z \) from the training data matrix \( X \) as the set of root nodes of the b-th tree, where \( b = 1, 2, 3, \ldots \), \( a \) and \( a \) are the number of root nodes, and the training data matrix \( X \) and sub sample matrix \( X_z \) are expressed by the formula:
   \[
   X = \begin{bmatrix} x_1, x_2, x_3, \ldots, x_n \end{bmatrix} \tag{4}
   \]
   \[
   X_z = \begin{bmatrix} x_{1z}, x_{2z}, x_{3z}, \ldots, x_{nz} \end{bmatrix} \tag{5}
   \]
   In formula (4, 5), the matrix is \( m \times n \) dimensional matrix, \( n \) is the number of variables, and \( m \) is the number of data samples contained in a single variable; \( u \) is the number of samples, \( u = mn \); \( X \) is \( m \times n \) dimensional matrix, \( n \) is the number of variables, \( m \) is the number of data samples contained in a single variable, \( 0 < m_z \leq m, 0 < n_z \leq n \).

2. **Step 2** perform binary segmentation on \( X_z \): randomly select a column vector \( x_j \), \( j \in \{1, 2, 3, \ldots, m_z\} \) from \( X_z \), and randomly select a cutting point \( T \) from the set of column vectors \( x_j \), as shown in formula (6).
   \[
   T = \min(x_j) + (\max(x_j) - \min(x_j))r \tag{6}
   \]
   In formula (6), \( r \) is a random number between 0 and 1. If \( X_z(i, j) < T \), all variables in row \( I \) of matrix \( X_z \) are divided into left subtree nodes. If \( X_z(i, j) \geq T \), all variables in row \( i \) of matrix \( X_z \) are divided into right subtree nodes, \( i = 1, 2, 3, \ldots, m_z \). When all the data in \( X_z \) is divided, the left subtree node set forms matrix \( X_{left} \) and the right subtree node set forms matrix \( X_{right} \).

3. **Step 3** record the path length \( h_n \) of the node where \( X_{left} \) and \( X_{right} \) are located. \( h_n \) is the number of edges from the root node to the current node. If \( h_n \) is greater than or equal to the tree height \( h_{max} \) or the number of sets in nodes is less than or equal to \( m_z \), stop training and complete the construction of a single isolation tree; otherwise, perform binary division for \( X_{left} \) and \( X_{right} \), and repeat step 3. The tree height \( h_{max} \) is:
   \[
   h_{max} = lbu \tag{7}
   \]

4. **Step 4** if \( b < a \), \( b = b + 1 \), repeat step 2 and step 3, otherwise stop training and complete the establishment of IF model. IF divides and isolates the data in \( X \). the normal data needs to be divided and isolated for many times and is in a high-density area; abnormal data needs to be isolated by a few divisions and located in low-density areas.

### 3.3. Eliminate abnormal data of power grid energy

After the data is calculated by the IF model, different high and low density areas are formed. The density area where the data is located is reflected by calculating the abnormal value score of the data,
and the data with high score is eliminated. The path length of $x_{ij}$ output from IF model is $h_{ij}$, and $x_{ij}$ is the element in matrix $X$. Calculate the outlier score of $x_{ij}$ through $h_{ij}$, and the formula is:

$$S(h_{ij}, u) = 2^\left(-E(h_{ij})\right) \left(\frac{C(u)}{u} - 1\right)$$  \hspace{1cm} (8)$$

$$C(u) = 2(\ln(u) + \xi) - \frac{2(u-1)}{u}$$  \hspace{1cm} (9)$$

In formulas (8 and 9), $c(u)$ is the average path length of all data in $X$; $\xi$ is Euler constant; $E(h_{ij})$ is the average path length of data $x_{ij}$ in an isolation tree. When the value of $S(h_{ij}, u)$ approaches 0.5, it indicates that the data has no obvious abnormal state; When the value of $S(h_{ij}, u)$ approaches 1, it indicates that the data is an abnormal value. After obtaining the outlier score of each data, combined with the equivalence and missing characteristics of the data, the following three kinds of data are eliminated from $X$: For the first data, $x_{ij}$ with outlier score $S(h_{ij}, u)$ greater than 0.75 in $X$ is eliminated and replaced by 0. For the second data, replace the missing data in $X$ with 0. For the third data, $c_1$ or more consecutive $x_{ij}$ with the same value in $X$ are eliminated and replaced with 0.

### 3.4. Complete the online cleaning of abnormal energy data of power grid

From the above analysis, it can be seen that the identification and cleaning of non-clean data is difficult to ensure the complete elimination of non-clean data, especially the multi-dimensional nature of power grid energy abnormal data, which brings difficulties to cleaning. With the gradual improvement of the fineness of the cleaning, the operation performance of the knowledge base system will gradually decrease, and the information contained in the abnormal energy data of the power grid will be gradually lost, which can not quickly and effectively solve the problems caused by the non-clean data. Data cleaning across data sources is mainly to ensure the correlation between metadata by matching the semantic constraints between multiple data sources. Assuming that the metadata feature set is $sdfrewa$, the following two constraints need to be met to complete metadata cleaning. First, if the two vertices of the edge representing the metadata relationship must be metadata objects, they are expressed by a formula, as shown in formula (10).

$$\forall e(f, t) \in E, f, t \in V$$  \hspace{1cm} (10)$$

In formula (10), $e$ refers to the directed edge of the metadata feature; $(f, t)$ refers to the coordinate point corresponding to the metadata feature; $E$ refers to the directed edge combination of metadata features; $V$ refers to the vertex set of metadata features. Second, if the relationship between two metadata objects can only be one-way, it can be expressed by formula, as shown in formula (11).

$$\forall e(f, t) \in E, e^{-1}(f, t) \notin E$$  \hspace{1cm} (11)$$

On the basis of satisfying formula (11), the syntax and semantics of energy metadata characteristics in the power grid operation process can be analyzed, and then the energy effective information can be extracted and tracked, so as to complete the online cleaning of power grid energy abnormal data. Therefore, through data cleaning, the quality of power grid energy data can be improved.

### 4. Experiment

#### 4.1. Experimental preparation

The grid energy data information of a cross-border power enterprise in recent 5 years is selected and introduced into the simulation experiment software to build a virtual power database. Two computers with Windows 10 and 16GB memory capacity are selected to simulate the sending and receiving of power grid energy data. Another computer with Windows 10 and 32GB memory capacity is selected to simulate the transmission environment of power grid energy data, and the experiment is carried out with the abnormal power grid energy data as the sample data. This paper uses the online cleaning method of power grid energy abnormal data based on improved random forest design to clean the
power grid energy abnormal data online, and then uses the traditional method to clean the power grid energy abnormal data online. 1000 grid energy anomaly data are randomly selected, of which 500 grid energy anomaly data are cleaned online by using the design method in this paper and set as the experimental group, and the other 500 grid energy anomaly data are cleaned online by using the traditional method and set as the control group. The root mean square error of the experimental group and the control group in the online cleaning of power grid energy anomaly data was compared. The lower the root mean square error, the higher the accuracy of the online cleaning method of power grid energy abnormal data, so as to obtain a higher accuracy online cleaning method of power grid energy abnormal data.

4.2. Experimental results and analysis
According to the above experiment preparation, complete the comparative experiment, record the data information generated during the experiment, and display the experimental results in the form of tables. The specific experimental results are shown in Table 1.

| Number | Data volume | Root mean square error of cleaning in the experimental group | Root mean square error of cleaning in the control group |
|--------|-------------|----------------------------------------------------------|-------------------------------------------------------|
| 1      | 50          | 0.1058                                                  | 0.8574                                                |
| 2      | 100         | 0.1414                                                  | 0.9978                                                |
| 3      | 150         | 0.1238                                                  | 0.7569                                                |
| 4      | 200         | 0.1824                                                  | 0.7581                                                |
| 5      | 250         | 0.1460                                                  | 0.8534                                                |
| 6      | 300         | 0.2052                                                  | 0.9580                                                |
| 7      | 350         | 0.1437                                                  | 0.9573                                                |
| 8      | 400         | 0.1841                                                  | 0.8581                                                |
| 9      | 450         | 0.1609                                                  | 0.9560                                                |
| 10     | 500         | 0.1513                                                  | 0.8577                                                |

According to Table 1, when the amount of grid energy abnormal data increases, the cleaning root mean square error of the experimental group is significantly lower than that of the control group, which can realize high-precision online cleaning of grid energy abnormal data. The main reason is that the experimental group can realize intermittent deletion mixed continuous deletion data cleaning based on the improved random forest. Therefore, the comparative experiments show that the online cleaning method proposed in this paper can effectively reduce the root mean square error of online cleaning of power grid energy anomaly data and improve the online cleaning accuracy. The mean square error of the cleaning data using this method is less than 0.21, while the mean square error of the cleaning data of the control group is higher than 0.75. It further ensures the quality of power grid energy data, meets the needs of power grid construction at the present stage, and has higher application value.

5. Conclusion
Aiming at the problem of energy abnormal data in the process of power grid operation, an online cleaning method of power grid energy abnormal data based on improved random forest is proposed in this paper. Firstly, the random forest algorithm is used to identify the outlier data. Then, outlier data, duplicate data and missing data are eliminated to obtain the mixed type missing data set. Finally, the improved random forest algorithm is used to compensate the missing data sets of mixed types. The effectiveness of IRF data cleaning algorithm is verified by taking the actual sewage data, and the following conclusions are obtained: the random forest algorithm can eliminate the outlier data in the power grid energy data. In the mixed type missing data set, the improved random forest algorithm has better cleaning effect than other algorithms, and is suitable for cleaning mixed type abnormal data. For data cleaning algorithms, some research institutions have proposed data preprocessing, sorting
neighbor method, multiple traversal data cleaning method, cleaning with domain knowledge, integrated data cleaning with database management system and so on. Based on the differences between Chinese data and Western data, Chinese data cleaning not only transplants the cleaning methods of Western data, but also has its own unique cleaning methods. According to the deficiency of current data cleaning research, the main research directions of data cleaning in the future are: the research and development of Chinese data cleaning tools; do in-depth research on the application of data mining methods in the field of data cleaning; the efficiency of duplicate record identification needs to be further improved; cleaning of unstructured data; interoperability between data cleaning tools; universality of data cleaning scheme. The improvement and optimization of Chinese data cleaning technology will be further studied in the follow-up.

References
[1] Govindarajan S, Ardila-Rey J A, (2021) Krithivasan K, et al. Development of Hypergraph Based Improved Random Forest Algorithm for Partial Discharge Pattern Classification[J]. IEEE Access, (99):1.
[2] Sihag P, Angelaki A, Jain B C. (2020) Estimation of the recharging rate of ground water using Random Forest Technique[J]. Applied Water Science, 10(7).
[3] Yao D, X Zhan, Kwoh C K. (2019) An improved random forest-based computational model for predicting novel miRNA-disease associations[J]. BMC Bioinformatics, 20.
[4] Asadi S, Roshan S E, Kattan M W. (2021) Random forest swarm optimization-based for heart diseases diagnosis[J]. Journal of Biomedical Informatics, 115(24):103-690.
[5] Su Y, K Weng, Lin C, et al. (2021) An Improved Random Forest Model for the Prediction of Dam Displacement[J]. IEEE Access, (99):1.
[6] Canion A, Mccloud L, Dobberfuhl D. (2019) Predictive modeling of elevated groundwater nitrate in a karstic spring-contributing area using random forests and regression-kriging[J]. Environmental Earth Sciences, 78(9).
[7] Xianjun, Liu, Xixiang, et al. (2019) Robust data cleaning methodology using online support vector regression for ultra-short baseline positioning system.[J]. The Review of scientific instruments, 90(12):124901-124901.
[8] Martinez-Luengo M, Shafiee M, Kolios A. (2019) Data management for structural integrity assessment of offshore wind turbine support structures: data cleansing and missing data imputation[J]. Ocean Engineering, 173(1):867-883.
[9] Ridzuan F, Wan M. (2019) A Review on Data Cleansing Methods for Big Data[J]. Procedia Computer Science, 161(3):731-738.
[10] Neira-Rodado D, Nugent C, Cleland I, et al. (2020) Evaluating the Impact of a Two-Stage Multivariate Data Cleansing Approach to Improve the Performance of Machine Learning Classifiers: A Case Study in Human Activity Recognition[J]. Sensors (Basel, Switzerland), 20(7).