A Line Loss Management Method Based on Improved Random Forest Algorithm in Distributed Generation System

Wang Zongbao

State Grid Baiyin Power Supply Company, Gansu Baiyin 730900, China
E-mail: 1136643177@qq.com

Received 14 May 2021; Accepted 27 May 2021;
Publication 27 August 2021

Abstract

The distributed power generation in Gansu Province is dominated by wind power and photovoltaic power. Most of these distributed power plants are located in underdeveloped areas. Due to the weak local consumption capacity, the distributed electricity is mainly sent and consumed outside. A key indicator that affects ultra-long-distance power transmission is line loss. This is an important indicator of the economic operation of the power system, and it also comprehensively reflects the planning, design, production and operation level of power companies. However, most of the current research on line loss is focused on ultra-high voltage (≧110 KV), and there is less involved in distributed power generation lines below 110 KV. In this study, 35 kV and 110 kV lines are taken as examples, combined with existing weather, equipment, operation, power outages and other data, we summarize and integrate an analysis table of line loss impact factors. Secondly, from the perspective of feature relevance and feature importance, we analyze the factors that affect line loss, and obtain data with higher feature relevance and feature importance ranking. In the experiment, these two factors are determined as the final line loss influence factor. Then, based on the conclusion of the line loss influencing factor, the optimized random forest regression algorithm is used to construct the line loss prediction model. The prediction verification results

Distributed Generation & Alternative Energy Journal, Vol. 37_1, 1–22.
doi: 10.13052/dgaej2156-3306.3711

show that the training set error is 0.021 and the test set error is 0.026. The prediction error of the training set and test set is only 0.005. The experimental results show that the optimized random forest algorithm can indeed analyze the line loss of 35 kV and 110 kV lines well, and can also explain the performance of 110-EaR1120 reasonably.

Keywords: Distributed power generation, 35 kV, line loss, random forest algorithm.

Introduction

Line loss is caused by converting part of the current into heat in the process of transmitting electric energy. Line loss is an important indicator of the economic operation of the power system, which comprehensively reflects the planning, design, production and operation level of the power enterprise. Power Supply Company A is located in the north-central part of Gansu Power Grid, and is responsible for the important task of power transmission from the west to the east and from the north to the south. At present, a large-scale regional power grid with 750 kV, 330 kV, and 220 kV as the main grid is formed. With the continuous expansion of the scale of the power grid in the jurisdiction, the grid structure changes relatively frequently, and the difficulty of managing the line loss of the main power grid is increasing. From the discovery of the abnormality of the line to the elimination, a lot of human resources are used, which takes a long time, and affects the economy and stability of the grid operation. In addition, the traditional monthly line loss statistics method for the same period cannot accurately reflect the actual abnormal root cause, and thus cannot effectively guide the development of line loss management. Moreover, in the past few decades, most of the research on line loss has focused on ultra-high voltage (≧110 KV) transmission lines, and less involved distributed power generation lines below 110 KV. In fact, the calculation of line loss is a very complex task, especially for low- and medium-voltage lines from 35 KV to 110 KV. The line is characterized by a large number of lines, a high load, a large amount of data, and the calculation is very complicated.

The theoretical calculation method of line loss can be roughly divided into two categories. One method is mathematical processing ideas based on the equivalent model. The limitation of these methods is that they ignore the impact of weather on the route. Another method is to perform autoregression based on historical data. The disadvantage of this method is that the prediction accuracy is very low and cannot meet the needs of the current power grid. With the continuous application of new digital technology systems in the power grid, the monitoring of medium and low voltage lines by power grid companies has become more and more perfect. There are more and more researches on line loss prediction and analysis of medium and low voltage lines using machine learning methods. At present, many scholars have done a lot of research on them. By evaluating the quality of line loss data, Wang et al. [1] proposed an evaluation model for evaluating the quality of line loss data in the same period based on the "rank sum" difference of penalty variable weight, which solved the problem of selecting the traditional long-term dependent index for line loss. Wang et al. [2] started from the perspective of the continuous improvement of the capacity of photovoltaic power plants integrated into the power grid, and made corresponding improvements to the IEEE14 node model, and analyzed the stability and influences when photovoltaic power plants were integrated into the rural distribution network. Liu et al. [3] proposed a data mining technology, which mainly extracted massive amounts of information from various data source systems accumulated by power companies, and built an anti-theft management system with resource sharing and decision support functions. This model solved the power load conditions such as line loss deviation, power difference, voltage and current imbalance. Li et al. [4] took machine learning as an entry point, and used neural networks to construct a relevant line loss model through a data-driven approach, and finally realized the accurate judgment of the location of the stealing. Xu et al. [5] studied how to use machine learning algorithms to build an abnormal line loss recognition model in the station area, and to realize the diagnosis of abnormal line loss in the power grid station area. In order to accurately calculate the daily line loss rate in the low-voltage transformer area, He et al. [6] proposed a multi-path network model with a denoising autoencoder to accurately evaluate the quality of the sampled data set and eliminate the line loss rate. Jin et al. [7] aimed at the problem of low accuracy of traditional anti-theft prediction methods, and proposed an anti-theft prediction method based on power big data. This method reconstructed the electricity theft data sample according to the abnormal rules. The experimental results showed that the prediction accuracy of this method was satisfactory, and it was efficient and feasible in the identification of stealing users.

The above-mentioned research has made major breakthroughs in the analysis of ultra-high voltage and power theft, but there is very little research on the medium and low voltage line loss of distributed power generation. Based on this, we take the 35 kV, 110 kV 1120 East Red Line (110-EaR1120) as the research object. Combining existing weather, equipment, operation, power outages and other data, we make an analysis table of line loss impact factors. From the two perspectives of line loss feature correlation and feature importance, we analyze the line loss influencing factors, and obtained data with higher feature correlation and feature importance ranking. Then, based on the conclusion of the line loss influencing factor, the optimized random forest regression algorithm is used to construct the line loss prediction model of the line.

1 Related Theories

1.1 Line Loss Calculation Theory

Line loss theory plays a very important role in loss reduction and energy saving, line loss management, etc. Through theoretical calculation of line loss, the distribution law of power loss in the line can be found, so that management and technical problems can be found. At present, the value of line loss is obtained by metering on the line. The specific calculation method is to subtract the value of the previous meter from the value of the latter. In this article, we use the theory of daily line loss. For lines greater than or equal to 35 kV, the line loss calculation is divided into two parts. The first part is the calculation of the loss of the components in operation such as transformers, lines, reactors, capacitors and main cameras. The essential idea is to use the root mean square current method (electric power method). Due to the different applications of new digital technologies in the power grid, each area is fully equipped with SCADA monitoring conditions, and the operation data within a day is also recorded and stored one by one. In order to reflect the impact of power generation and load changes on line loss to the greatest possible extent, this article assumes that the power output and load per hour remain unchanged. The calculation formula of daily power loss is shown in Equation (1), and the calculation formula of daily line loss rate is shown in Equation (2) [8]:

$$\Delta A_d = \left[\sum_{i=1}^{n} \left(3 \sum_{t=1}^{24} I_{ti}^2 R_{ti} \right) + \sum_{j=1}^{k} \sum_{t=1}^{24} V_{tj}^2 G_{mj} \right] * 10^{-3} \tag{1}$$

$$\Delta A_d\% = (\Delta A_d / A_d) * 100\% \tag{2}$$

Where: ΔA_d represents the daily power loss (kWh). n represents the number of power grid lines and transformer branches. k represents the number of power grid transformers and the number of branches grounded by other components. R_{ti} represents the resistance of the $j - th$ branch at time t, and it is basically a constant when temperature changes are not considered. G_{mj} represents the conductance of each element in the $j - th$ transformer. I_{ti} represents the current (A) of the $j - th$ branch at time t. V_{tj} represents the voltage (V) of the $j - th$ node at time t. I_{ti} and V_{tj} are obtained by solving the nodal and power equations by Newton's method.

1.2 Improved Random Forest Algorithm

1.2.1 Principles of random forest algorithm

Random forest is one of the classic algorithm models of machine learning. In the past ten years, random forests have developed rapidly, especially in the field of bioinformatics [9, 10], economic management [11], medical field [12], criminal investigation field [13] and pattern recognition field. The shortcoming of the random forest algorithm is in the classification of the data. Therefore, many scholars have made a lot of improvements to the algorithm. For example, Huang et al. [14] compared the performance of random forest and support vector machine in processing unbalanced data, and found that they are more sensitive to unbalanced classification data. But at present, there are relatively few researches on the improvement of random forest. Therefore, it is meaningful to optimize and improve the random forest and then apply it to the line loss calculation of medium and low voltage lines [15].

Random forest is mainly realized through Bootstrap technology. In addition, there are many reports on the security protection of smart grids [17-20]. The first step is to randomly select *k* samples from the original training sample set *N* to generate a new training sample set [16]. The second step is to generate *k* decision trees based on the sample set, and randomly combine them to obtain a random forest [21, 22]. It is worth noting that the classification result of the new data is determined by the number of votes formed by the decision tree. *D* is the sample set, D_1, D_2, and D_k are the decision trees generated after each random sampling [23, 24]. The schematic diagram of random forest is shown in Figure 1.

The essence of the random forest algorithm is to arrange and combine multiple decision trees. The establishment of each tree only relies on an independent sample. The basic method of the random forest algorithm is to split each node by a random method and compare the errors generated in different situations. In general, the more decision trees there are, the greater the probability of getting a better classification effect. After *k* rounds of

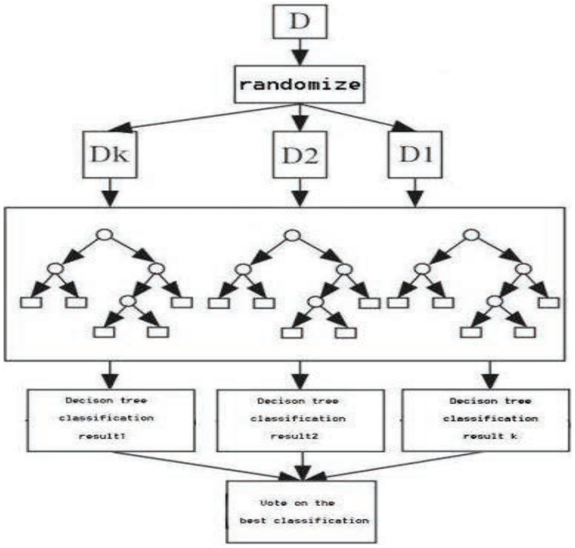
D
randomize
Dk
D2
D1
Decison tree classification result1
Decison tree classification result2
Decison tree classification result k
Vote on the best classification

Figure 1 The working principle of random forest.

training, a sequence $\{h_1(X), h_2(X), h_3(X), \ldots, h_k(X)\}$ is obtained, and after voting, the final decision is shown in Equation (3):

$$H(x) = \arg\max_{Y} \sum_{i=1}^{k} I(h_i(x) = Y) \tag{3}$$

Among them, $H(x)$ represents the combined classification model, h_i represents the classification result of a single decision tree, Y represents the output target variable, and I is the indicative function.

1.2.2 Improved random forest algorithm

The monitoring data of medium and low voltage lines are characterized by multiple dimensions and complex types. One problem that is often encountered is serious lack of data. Therefore, before feeding a large amount of data to the random forest algorithm, the data needs to be preprocessed. The improvement of the random forest algorithm in this paper is mainly to add the weights of multiple influence factors in the process of randomization to generate the decision tree. The calculation process of the correlation coefficient in the specific weighting process is shown in Equation (4), and the calculation process of the correlation degree is shown in Equation (5). Finally, we assign values to D_1, D_2, and D_k respectively to get a new vector. The purpose is to embody key influencing factors in the process of randomization and weaken non-key influencing factors.

$$\varsigma(k) = \frac{\min_i \min_k |x_0(k) - x_i(k)| + \rho * \max_i \max_k |x_0(k) - x_i(k)|}{|x_0(k) - x_i(k)| + \rho * \max_i \max_k |x_0(k) - x_i(k)|}$$

$$k = 1, \ldots, m \tag{4}$$

$$r_i = \frac{1}{m} \sum_{k=1}^{m} \zeta_i(k) \tag{5}$$

In the formula, ζ is the correlation coefficient, ρ is the resolution coefficient, and the value range of ρ is between (0, 1). The smaller ρ indicates the greater the difference between the correlation coefficients. Generally, the value of ρ is 0.5, and ri represents the degree of correlation.

In the design process, in addition to considering the correlation between the factors that affect the data during randomization, the correlation between the various decision trees is also considered. The correlation degree here uses the mean value method, that is, in order to prevent the difference from being too large, Equation (6) is used here to make a second assignment of the correlation degree for each decision tree. In this way, each decision tree will get a result, and finally these results will be coupled to get the final result. The improved algorithm structure is shown in Figure 2.

$$\zeta_h = \frac{\sum_{i=1}^{k} \zeta_{hi}}{\sum_{i,j=1}^{k} \zeta_{ij}} \tag{6}$$

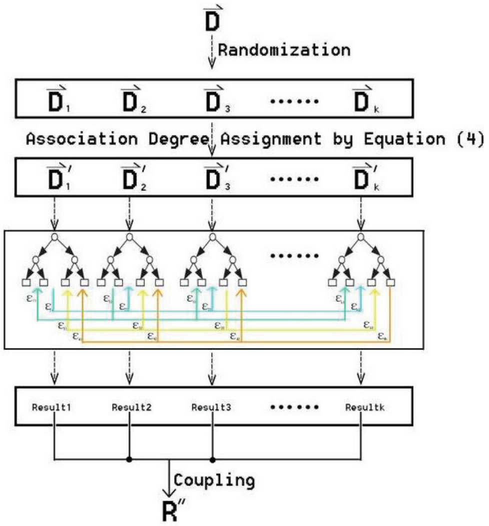
$\vec{D}$
Randomization
$\vec{D}_1$ $\vec{D}_2$ $\vec{D}_3$ ...... $\vec{D}_k$
Association Degree Assignment by Equation (4)
$\vec{D}'_1$ $\vec{D}'_2$ $\vec{D}'_3$ ...... $\vec{D}'_k$
Result1 Result2 Result3 ...... Resultk
Coupling
$\vec{R}''$

Figure 2 Random forest algorithm with correlation degree assignment.

2 Experiment

2.1 Data Preparation

This experiment selects 35 kv and 110 kv EaR1120. The data sources are PMS2.0 system, power consumption information collection system, integrated power and line loss management system, power transmission and transformation online monitoring system and corresponding historical weather data. It mainly includes model parameters and meter values of 35 kV and 110 kV lines. The processing flow of all data is shown in Figure 3. When the extracted data is abnormal data, we use smoothing numerical method to fill it. The data obtained from the power system has some missing data. After statistical analysis, it is found that the missing rate is 0.8%. This shows that these missing data will not have a great impact on the original data, so this paper also uses smooth numerical methods to fill them.

After all the key data is collected, in order to ensure the quality of the data to meet the needs of data analysis and model construction, our first job is to preprocess the data. The quality of data such as line model parameters in the integrated power and line loss management system is relatively high. However, due to meter failure, terminal disconnection, etc., many meter data will be missing or counted abnormally. That leads to abnormal conditions

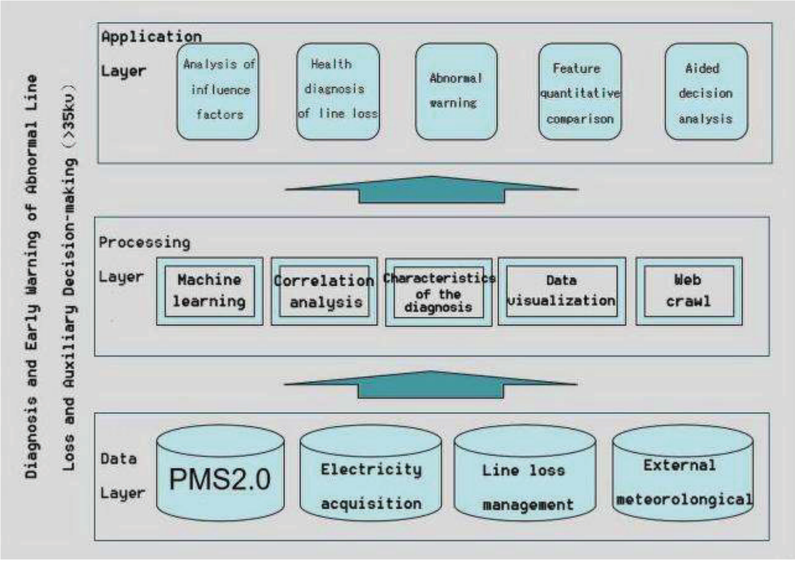

Figure 3 The overall data processing flow.

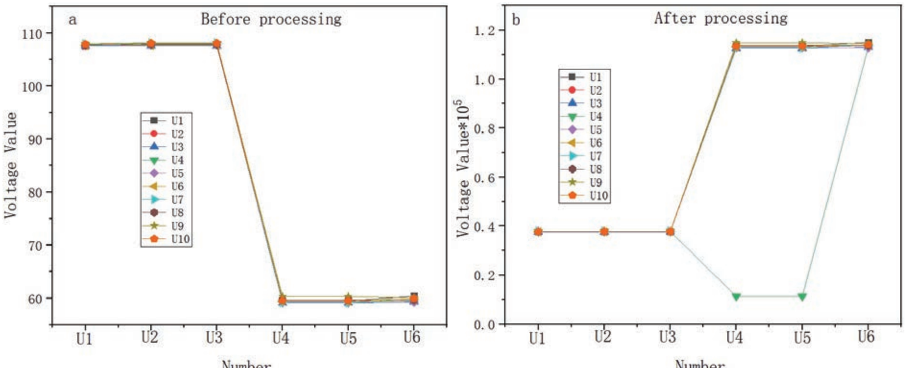

Figure 4 Voltage distribution diagram before and after treatment.

such as null and 100% line loss rate. We filter out such data directly. We replace the zero value of the voltage with a null value, and use the smoothing value method to fill in the missing value of the voltage. And according to the meter connection mode(three-phase three-wire system, three-phase four-wire system), transformer transformation ratio and other data are converted into line voltage, we unify the date format into numerical data, as shown in Figure 4. The weather data is complete and of high quality, but the format of each field is not uniform and cannot be used directly. We formulate different cleaning strategies according to the respective characteristics of each field, and convert the original data into numbers to facilitate the next calculation.

2.2 Synthesizing a New Data Set

When the data preparation is complete, the data needs to be converted into a new data set that meets the requirements of the improved random forest algorithm. The main reserved key fields in the new data set include component number, measuring point number, component name, date, etc. At the same time, we use data conversion to generate new variables, as shown in Table 1.

After the data is converted, the dimension difference between the obtained data is large. In order to eliminate the influence of the dimension on the analysis results, we used the normalization method to eliminate the data with a larger dimension, as shown in Figure 5. The normalized data is obviously much smaller than the original data, which greatly reduces the difficulty of data analysis.

Table 1 Generating new variable name

Before the Variable Name Changed	After the Variable Name Changed
A-phase voltage	A-phase average voltage
B-phase voltage	B-phase average voltage
C-phase voltage	C-phase average voltage
Active power	Average active power
Reactive power	Average reactive power
Daytime and nighttime temperature	Average temperature
Daytime and nighttime humidity	Average humidity
Day and night wind	Average wind
Daytime and nighttime precipitation	Average precipitation

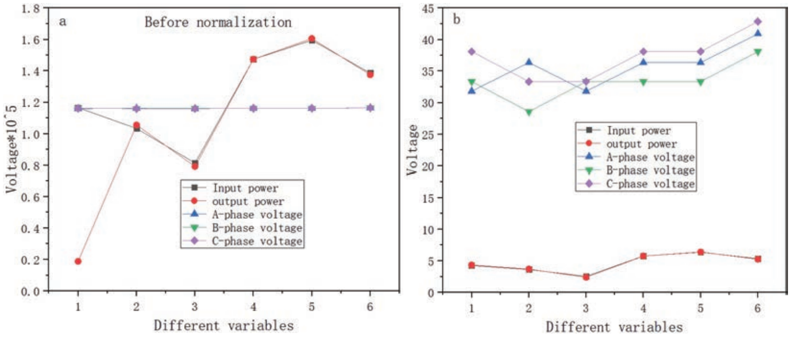

Figure 5 Data before and after normalization.

Finally, we get a new data set. The key fields included in the new data set are line component number, metering point number, date, line loss, electricity, voltage, active power, reactive power, and meteorological data. Part of the data in the new data set is shown in Figure 6.

2.3 Experimental Results of the Improved Random Forest Algorithm

In the experiment, the new data set is fed into the improved random forest algorithm model. The first step is to analyze the current line loss status of EaR1120 through the model, the second step is to analyze the key influencing factors of the line loss, and the third step is to diagnose the health status of the line. Finally, the adjustment strategy of the influence factor of line loss is given.

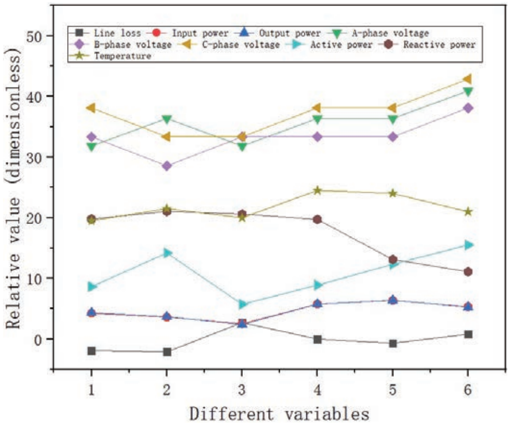

Figure 6 Key fields in the new data set.

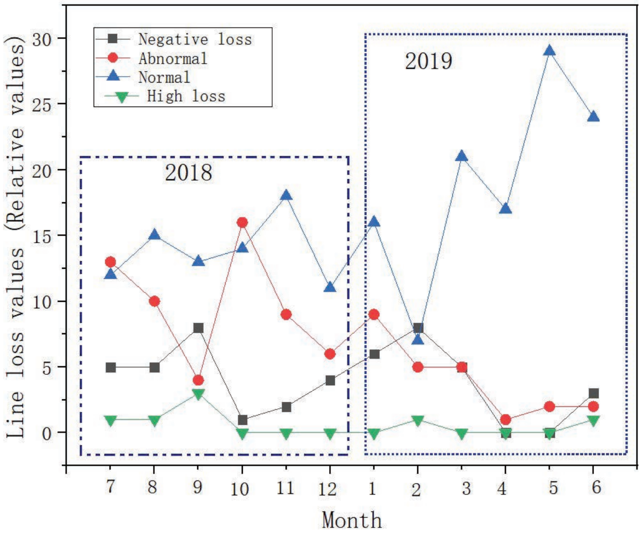

Figure 7 Monthly line loss of EaR1120.

2.3.1 Analysis of the current status of line loss

We first use the improved random forest algorithm model to visually analyze the new data set, and the analysis results are shown in Figure 7. The data cycle of the line loss data of EaR1120 is from July 2018 to June 2019, a total of 365 days. After removing missing values and abnormal value (line loss is 100%), the effective line loss days are 333 days. Among them, the number of negative loss days for the line was 47 days, accounting for 14.11%. The number of abnormal days of line loss was 82 days, accounting for 24.62%. The normal number of days for line loss is 197 days, accounting for 59.16%. The number of high-loss days on the line is 7 days, accounting for 2.10%. The number of days when the line loss was negative was 129 days, accounting for 38.74%. At the same time, Figure 7 shows that from July 2018 to June 2019, the line loss of EaR1120 shows the overall characteristics of relatively stable in the early period, volatility in the mid-term, and improvement in the later period. However, there is a rebound trend in June 2019, and follow-up line loss changes need to be paid attention to.

2.3.2 The importance analysis of influencing factors

In this part of the experiment, we first did an autocorrelation analysis on the factors that affect the line loss. Many experiments have shown that the component number, metering point number, etc. obviously have no effect on the line loss. The relationship between other factors is shown in Table 2. As can be seen from Table 2, the correlation coefficient between input power and output power is 0.99. The correlation coefficients among A-phase voltage, B-phase voltage, and C-phase voltage all exceed 0.9. Therefore, feature autocorrelation processing is required. Our approach is to delete the output power, B-phase voltage, and C-phase voltage, and retain the input power and A-phase voltage. Finally, the retained parameters are substituted into the subsequent model.

After deleting the impact factors with high autocorrelation, we substitute the data into the improved random forest model again. After proper adjustment of the parameters, we calculate the feature importance of each influencing factor. The calculation results are shown in Table 3.

According to the feature importance in Table 3, we choose the influencing factor whose importance accounted for 85% as the key influencing factor of line loss. At the same time, we define that when the importance of the dependent variable exceeds 85%, it will be characterized as the determinant of the line loss. Through this determination method, we discarded other less important impact factors. Finally, the input power, active power, A-phase voltage, and reactive power are determined as the main line loss influencing factors of EaR112.

2.3.3 Diagnosis of line damage health situation

After determining the important influence factor of the line loss through the improved random forest algorithm, we begin to design the entire line loss

Table 2 Correlation between line loss influencing factors

Field	Input Power	Output Power	A-phase Voltage	B-phase Voltage	C-phase Voltage	Active Power	Reactive Power	Temperature	Humidity	Wind	Precipitation
Input power	1	0.99	−0.13	−0.12	−0.06	0.6	−0.03	−0.04	−0.13	−0.01	−0.16
Output power	0.99	1	−0.14	−0.13	−0.06	0.6	−0.03	−0.04	−0.13	−0.02	−0.17
A-phase voltage	−0.13	−0.14	1	0.91	0.95	−0.18	−0.65	0	0.09	−0.06	0.01
B-phase voltage	−0.12	−0.13	0.91	1	0.91	−0.19	−0.62	−0.1	0.06	−0.09	−0.01
C-phase voltage	−0.06	−0.06	0.95	0.91	1	−0.11	−0.71	0.01	0.12	−0.03	0.01
Active power	0.6	0.6	−0.18	−0.19	−0.11	1	−0.03	0.01	−0.14	0.01	−0.15
Reactive power	−0.03	−0.03	−0.65	−0.62	−0.71	−0.03	1	−0.21	−0.15	−0.04	−0.08
Temperature	−0.04	−0.04	0	−0.1	0.01	0.01	−0.21	1	0.1	0.28	0.27
Humidity	−0.13	−0.13	0.09	0.06	0.12	−0.14	−0.15	0.1	1	0.06	0.55
Wind	−0.01	−0.02	−0.06	−0.09	−0.03	0.01	−0.04	0.28	0.06	1	0.22
Precipitation	−0.16	−0.17	0.01	−0.01	0.01	−0.15	−0.08	0.27	0.55	0.22	1

Table 3 The importance of features of each impact factor

Field	Importance
Input power	0.3600
Active power	0.2376
Phase A voltage	0.1406
Reactive power	0.1328
temperature	0.0556
humidity	0.0403
Field	importance
Input power	0.3600
Active power	0.2376

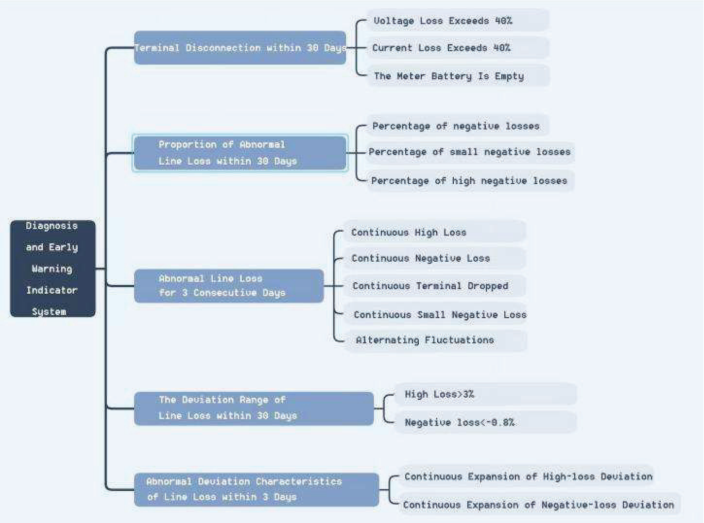
Diagnosis and Early Warning Indicator System
Terminal Disconnection within 30 Days
Voltage Loss Exceeds 40%
Current Loss Exceeds 40%
The Meter Battery Is Empty
Proportion of Abnormal Line Loss within 30 Days
Percentage of negative losses
Percentage of small negative losses
Percentage of high negative losses
Abnormal Line Loss for 3 Consecutive Days
Continuous High Loss
Continuous Negative Loss
Continuous Terminal Dropped
Continuous Small Negative Loss
Alternating Fluctuations
The Deviation Range of Line Loss within 30 Days
High Loss>3%
Negative loss<-0.8%
Abnormal Deviation Characteristics of Line Loss within 3 Days
Continuous Expansion of High-loss Deviation
Continuous Expansion of Negative-loss Deviation

Figure 8 Line loss diagnosis and early warning indicator system.

diagnosis system. Combining expert opinions and business reality, we have formulated a comprehensive line loss diagnosis and early warning indicator system, which is implemented by implementing a deduction system of 100 points, as shown in Figure 8.

After the construction of the indicator system is completed, we first diagnose the disconnection of the mid-term terminal. The specific definition is that the three situations where the voltage missing item exceeds 40%, the current missing item exceeds 40%, and the power at the bottom of the

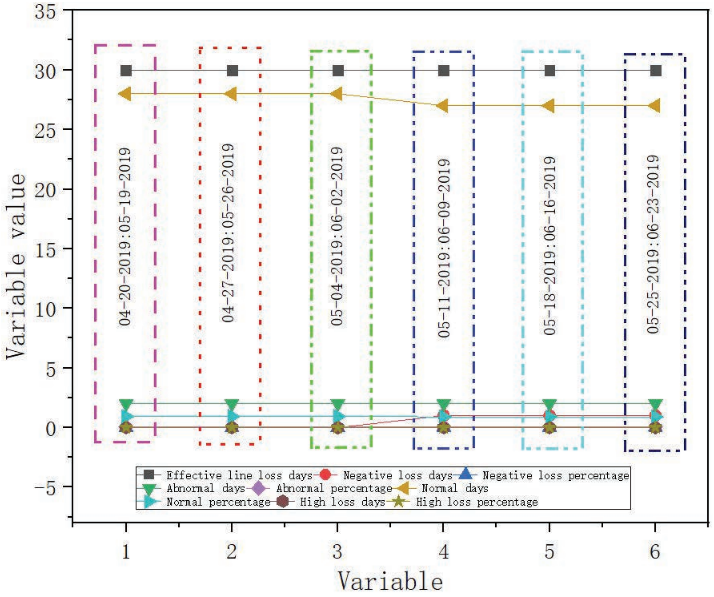

Figure 9 Medium- and short-term line loss of EaR1120.

meter are empty on the day are regarded as terminal offline events. We judge the collection of line power consumption information in the last 30 days, and calculate the percentages of missing fields in the collection of voltage, current, and power at the bottom of the meter. Finally, we count the dropped calls in the past 30 days. The specific scoring rules are as follows: 0.5 points will be deducted for each day of disconnection, and 15 points will be deducted for all disconnections within 30 days. According to this rule, the mid-term terminal drop score of this line is diagnosed.

Immediately after that, we diagnosed the abnormal situation of the short-term line loss. The lines in the last 30 days are classified according to the line loss type. We respectively calculate the number of days and the proportion of line loss under the three abnormal conditions of negative loss, high loss, and small loss. If the sum of the proportions of the three types is less than 30%, no points will be deducted. If it exceeds 30%, then 1 point will be deducted for every 1%, until 20 points are deducted. Finally, the mid-term line loss health status of the line is obtained. In the experiment, the line loss situation of the last 3 days is analyzed. We focus on judging whether there are five situations such as continuous negative loss, high loss, small loss, terminal drop, and alternating fluctuations. The scoring rules are as follows: no points will be deducted if none of the 5 situations occurs. 25 points are deducted for consecutive high or negative losses. 20 points will be deducted for continuous terminal disconnection. 15 points are deducted for consecutive small losses. 10 points deducted for alternating fluctuations. Finally, the short-term line loss health status of the line is obtained. Taking the line loss of EaR1120 as an example, the calculation result is shown in Figure 9. It can be seen that, as time goes by, the number of days with high or negative line loss will continue to increase by the end of June.

2.3.4 Adjustment strategy of line loss

According to the main influencing factors of EaR1120 obtained above, we have carried out a detailed analysis of EaR1120 from the four aspects of voltage, active power, reactive power and daily electricity.

It can be seen from Figures 10a, 10b, and 10c that the law of the three-phase voltage is basically the same. When the line loss is normal, the operating voltage fluctuation range of EaR1120 is approximately 115750–116250 V. When the line loss is abnormal, the voltage is basically the same as when the line loss is normal. When the line loss is negative and high, the voltage is higher than the voltage when the line loss is normal. When the line is negative, it is about 350 V high. When the line is high loss, it is about 1000 V high. It can be seen from Figure 10d that when the line loss is high or negative, the active power of the line has a big gap compared to the normal. The line loss is about 50% of the normal line loss. When the line loss is high, it is about 30% of the normal line loss. When the line loss is abnormal, the active power is slightly higher than when the line loss is normal.

It can be seen from Figure 11a that when the line loss is abnormal, the line reactive power is lower than the normal reactive power. Among them, when the line is high loss, the low amplitude is the largest, and the reactive power is about 50% of the normal time. When the line is negatively damaged, the reactive power is about 80–90% of normal. When the line loss is abnormal, the low amplitude is the smallest, and the reactive power is about 90% of the normal. It can be concluded from Figure 11a that, within a certain range of line loss, the line loss is inversely proportional to the daily power. That is, with the increase of line power, the range of negative or high line loss is gradually narrowing. When the daily power of the line exceeds 400,000kWh, there will be basically no large line loss.

Based on the above influencing factors, it can be seen that when the line loss of EaR1120 is normal, the load is large and the voltage drops. In order to maintain voltage stability, reactive power compensation is increased, so the

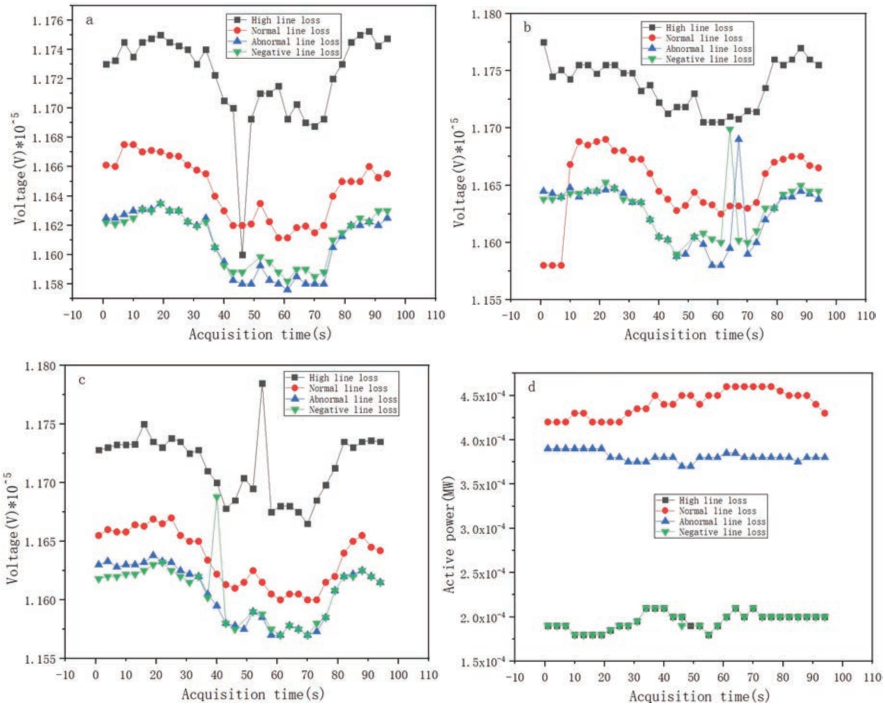

Figure 10 (a) Comparison of voltage characteristics of A-phase at EaR112000018333476, (b) Comparison of voltage characteristics of B-phase at EaR1120 00018333476, (c) Comparison of voltage characteristics of C-phase at EaR1120 00018333476, (d) Comparison of the average active power at EaR1120 00018333476.

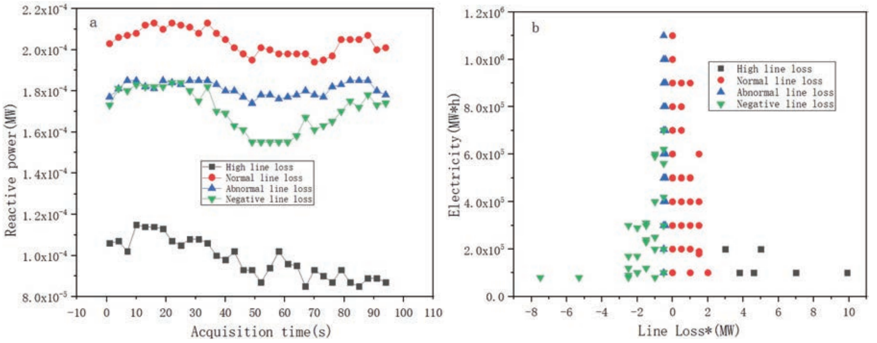

Figure 11 Reactive power and daily electricity.

reactive power is maximum when the line loss is normal. When the line loss is abnormal, the load is small and there is no obvious voltage drop, so the reactive power is smaller than when the line loss is normal.

3 Conclusion

In short, for the line loss problem of medium and low voltage lines, this paper proposes an improved random forest algorithm. From the perspective of line loss feature correlation and feature importance, the line loss impact factors are analyzed separately, and these two factors are determined as the final line loss impact factors. Then, based on the conclusion of the line loss influencing factor, the optimized random forest regression algorithm is used to construct the line loss prediction model of the line. The prediction and verification results show that the line loss condition of EaR1120 obtained by simulation is completely consistent with the actual situation, and the performance of EaR1120 can be explained reasonably. In the future, we will extend the application of this model to circuits above 110kv for application. In the application process, we will continue to improve the algorithm model and strengthen the generalization ability of the algorithm model.

References

[1] F. Wang, W. Liu, X. Chen, W. Wang, Z. Xing. Evaluation Model of Synchronous Line Loss Data Quality Based on Penalty Variable Weight RDA. Electric Power, 2020, 53(12): 223–231.

[2] J. Wang, Y. Wang, Y. Yang, Y. Wang. Analysis the Impact of Photovoltaic Power Station's Integration into Rural Distribution Network. Electric Power, 2018, 51(6): 150–154.

[3] L. Liu, C. Zhu. Application of Data Mining Technology to Build the Anti-Stealing Management System. Electric Power, 2017, 50(10): 181–184.

[4] Z. Li, H. Hou, Y. Jiang, C. Wan, R. Zheng. Line Loss Calculation and Electricity Theft Analysis Based on Artificial Neural Network. Southern Power System Technology, 2019, 13(2): 7–12.

[5] Y. Xu, H. Shi. Research on Diagnosis Method of Line Loss Abnormality in Power Grid Based on Machine Learning. Electric Power, 2020, 11(1): 236–237.

[6] W. He, Y. Sun, J. Jiang, L. Jin. Reference Test of Daily Line Loss Rate in Low Voltage Transformer Area based on RNN. Computer Measurement & Control, 2020, 28(9): 58–64.

[7] B. Jin, M. Zhang, H. Wu, Y. Shi. A prediction method of anti-electricity stealing based on big data of electric power. Journal of Light Industry, 2020, 95(4): 81–87.

[8] L. Tao, S. Liu, X. Ceng. Theoretical calculation and analysis of line loss based on power grid. Technology Innovation and Application, 2018, 23(4): 132–133.

[9] X. Chen, M. Liu. Prediction of protein-protein interactions using random decision forest framework. Bioinformatics, 2005, 21(24): 4394–4400.

[10] Smith A, Sterba-Boatwright B, Mott J. Novel application of a statistical technique, Random Forests, in a bacterial source tracking study. Water Research, 2010, 44(14): 4067–4076.

[11] Y. Cheng, H. Zou. Random Forest RFM Model and Its Evaluation in Bank Credit Risk. Journal of Anqing Teachers College (Natural Science Edition), 2018, 3(1): 34–37.

[12] T. Sun, X. Xu. Medical Big Data Analysis and Clinical Application Based on Machine Learning. Software Guide, 2019, 11(1): 10–14.

[13] R. Lu, L. Li. Model of Crime Prediction Based on the Random Forest. Journal of Criminal Investigation Police University of China, 2019, (3): 108–112.

[14] Y. Huang, W. Cha. Comparison on Classification Performance Between Random Forests and Support Vector Machine. Software, 2012, 33(6): 107–110.

[15] J. Zhao, X. Zhang, F. Di, S. Guo, X. Li. Exploring the Optimum Proactive Defense Strategy for the Power Systems from an Attack Perspective. Security and Communication Networks, 2021, 2021(1): 1–14.

[16] Y. Tao, T. Huang, M. Li, et al. Research on Log Audit Analysis Model of Cyberspace Security Classified Protection Driven by Knowledge Map[J]. Netinfo Security, 2020, 20(1): 46–51.

[17] W. Dong, Y. Li. Research on Analysis of Attacks on Smart Grid Network Based on Complex Network[J]. Netinfo Security, 2020, 20(1): 52–60.

[18] W. Luo, C. Xu. Network Intrusion Detection Based on Improved MajorClust Clustering[J]. Netinfo Security, 2020, 20(2): 14–21.

[19] C. Peng, Y. Zhao, M. Fan. A Differential Private Data Publishing Algorithm via Principal Component Analysis Based on Maximum Information Coefficient[J]. Netinfo Security, 2020, 20(2): 37–48.

[20] R. Wang, C. Ma, P. Wu. An Intrusion Detection Method Based on Federated Learning and Convolutional Neural Network[J]. Netinfo Security, 2020, 20(4): 47–54.

[21] J. Xiong, R. Bi, M. Zhao, J. Guo, Q. Yang. Edge-assisted privacy-preserving raw data sharing framework for connected autonomous vehicles, IEEE Wireless Communications, 2020, 27(3): 24–30.

[22] Y. Tian, Z. Wang, J. Xiong, J. Ma. A blockchain-based secure key management scheme with trustworthiness in DWSNs, IEEE Transactions on Industrial Informatics, 2020, 16(9): 6193–6202.

[23] J. Xiong, X. Chen, Q. Yang, L. Chen, Z. Yao. A task-oriented user selection incentive mechanism in edge-aided mobile crowdsensing, IEEE Transactions on Network Science and Engineering, 2020, 7(4): 2347–2360.

[24] J. Xiong, R. Ma, L. Chen, Y. Tian, Q. Li, X. Liu, Z. Yao. A personalized privacy protection framework for mobile crowdsensing in IIoT, IEEE Transactions on Industrial Informatics, 2020, 16(6): 4231–4241.

Biography

Wang Zongbao, born in June 1991, graduated from Northeast Dianli University, majoring in electrical engineering. The current deputy dispatcher of the Power Dispatching Control Center of State Grid Gansu Electric Power Company State Grid Baiyin Power Supply Company, mainly engaged in power grid economic dispatch, power grid security and stability analysis, and big data applications.Successively presided over the compilation of “Baiyin Power Grid Monitoring Information Management System” and “Concurrent Line Loss Offline Auxiliary Calculation and Management System”. In 2018, the QC project “Research and Development and Application of Auxiliary

Calculation and Management System for Line Losses in the Same Period" won the second prize of Excellent QC Achievement of Baiyin Power Supply Company. In 2018, he presided over the "Big Data-Based Diagnosis and Decision-making of Abnormal Line Losses in the Same Time" project, which won the gold prize of the Gansu Provincial Company Data Value Mining Innovation Competition. In 2019, he hosted the "Big Data-based Line Loss Line Abnormal Diagnosis and Decision-making in the Same Time", and obtained the software copyright of "35 kV and above Line Line Loss Analysis Tool Software".