Missing data filling method based on linear interpolation and LightGBM

Guilin Huang

Electrical and Electronic Engineering, Imperial College London,
London, SW7 2AZ, United Kingdom
gh1116@ic.ac.uk

Abstract: In the context of the rapid development of integrated energy and the digital transformation of power grids, data is playing an increasingly important role in the safe operation of power grids. To deepen the value of data application and ensure the accuracy of data application, this paper proposes a data filling method that combines linear interpolation and LightGBM (Light Gradient Boosting Machine) in response to the missing phenomenon in the source network data collection process. The process can generally be divided into 2 steps: First, linear interpolation is exploited to process short-term missing data. Then the LightGBM can be used to process long-term missing data. In the process of using LightGBM, linear interpolation is used to interpolate the independent variables of the input model. Through the above process, the data for the missing ratio could be obtained, which can then be used to complete all data filling in order from high to low. Through actual data test, this method has better data filling performance.

1. Introduction
At the end of 2019, the installed capacity of new energy generation in the operation area of State Grid Corporation of China was 350 million kilowatts, accounting for more than 20% of the total installed capacity. It is estimated that by 2050, the proportion of national wind power and photovoltaic power generation capacity will reach 24% and 31% respectively. The high proportion of new energy access has brought a great impact on the safe and stable operation of the power grid. To improve the regulation level of the power system and increase the anti-interference ability of the power system, the Shandong network source supervision service technology platform centred on regional grid coordination services, focusing on the pain points and demand orientation of power supply and power grid business, through data driving, mining the value of data, optimizing power production methods, it plays an increasingly prominent role in supporting the stable and safe operation of large power grids.

In the field of data mining, data preprocessing plays a very important role. Data mining for correct and complete data can obtain valuable information, which in turn allows users to make correct decisions [1]. However, there will inevitably be a variety of dirty data in the data set, including duplicate records, missing values and wrong data, which will affect the correctness of data mining algorithms certainty. Among them, the key problem is the missing value problem, and many scholars have done a lot of research on it. After decades of development, the research on data missing value filling has a relatively mature theoretical basis and algorithm [2-4]. In summary, the filling model of data missing value can be divided into two categories: the method based on the parametric model and the method based on the non-parametric model. In terms of the method based on parameter model,
Chen S [5] in 2014 proposed a missing value filling algorithm based on quantile regression, which combines quantile regression algorithm with general empirical likelihood estimation method to estimate Missing value; Shah AD [6] analyzed that in the case of random missing, the parameter filling algorithm based on the random forest is used to fill the non-linear missing value problem, which proves that the random forest filling strategy is very good in nonlinear spatial data. Fill effect. Based on the method of the non-parametric model, Masseran.N [7] compared the effects of several different single filling methods, and discussed the impact of different missing modes on filling; Madhu [8] proposed a method for continuous attribute data The non-parametric discretization filling method is used to fill the missing

Based on the above research methods, to take advantage of the integrated model, this article combination linear interpolation and LightGBM to fill in the missing values of the data set, and through an example test, this method has a better effect on data filling

2. Algorithm theory

2.1 Linear interpolation

Linear interpolation is an interpolation method for one-dimensional data. It estimates the data value based on the two data points adjacent to the point that needs to be interpolated in the one-dimensional data sequence. Among them, the simplest is the first-order linear interpolation, as follows The picture shows:

![Figure 1 Schematic diagram of linear interpolation](image)

The above figure shows the principle of linear interpolation. In practical applications, according to the distribution of data, low-order to high-order linear interpolation is selected.

2.2 LightGBM

LightGBM is an open-source, fast and efficient decision tree algorithm-based enhancement (GBDT, GBRT, GBM and MART) framework announced by Microsoft Research Asia in 2017 [9]. It is used for sorting, classification, regression and other machine learning. Support high-efficiency parallel training. The decision tree sub-model in LightGBM uses the method of splitting nodes by leaves. Therefore, its computational cost is relatively small. It is precisely because this split method is selected that the depth of the tree and the minimum of each leaf node needs to be controlled. The amount of data to avoid over-fitting. LightGBM chose a decision tree algorithm based on Histogram to divide the feature value into a lot of small "buckets", and then look for splits on these "buckets", which can reduce storage costs and computational costs. Besides, the processing of category features also enables LightGBM to have a better improvement under certain data.

3. Missing data filling method based on linear interpolation and LightGBM algorithm

When filling missing data, first consider the logical relationship between data features for filling, then consider establishing mathematical model analysis, and finally consider generating models to ensure that the data filling program is simple and reliable. Therefore, because of the correlation between the
missing amount of the filled data and the characteristic data, the method of combining linear interpolation and LightGBM is used to fill the missing values. The specific steps are as follows:

1. For multi-dimensional data to be filled, according to the number of consecutive missing data, it will be divided into two types of missing types. The missing data is less than \( \delta \) as missing type A, and the others as missing type B;
2. Use linear interpolation to fill in the missing data of type A at each measuring point;
3. After linear interpolation, count the missing data of each measuring point, and take the measuring point with the largest missing quantity as the priority filling point \( y \), and other measuring points as the model input quantity \( x \);
4. For the measuring point \( x \) in step 3, for measuring points with missing data, first use linear interpolation to preprocess the data of all vacant positions to improve the accuracy of the model;
5. Use LightGBM to fill the measuring point \( y \) in step 3;
6. Repeat steps 3 to 4 until all the data is filled. The specific process is as follows:

---

Figure 2 Algorithm design flow chart
4. Case Analysis

4.1 Data set
This paper selects the relevant data in the gate control process of a thermal power unit as the experimental data. To facilitate error verification, the experimental data is randomly discarded to construct a data missing test set, as follows:

| Sample length | Measuring point | The proportion of missing measuring point data | Short-term missing threshold δ |
|---------------|----------------|-----------------------------------------------|-------------------------------|
| 80000         | Main control command, Main steam pressure, Active power, Opening degree of regulating door 1 | 0.26~0.39 | 10 |

4.2 Experimental results
In order to verify the performance of the algorithm in this paper, for the same data set, the random forest method is selected for comparative analysis, and the following experimental results are obtained.

| Column name | Maximum error | Minimum error | Mean error | Missing ratio | Time used (s) |
|-------------|---------------|---------------|------------|---------------|---------------|
| Main control command | 0.028855 | 3.52E-08 | 0.003056 | 0.352163 | 5.770381 |
| Main steam pressure | 0.067292 | 1.16E-07 | 0.004838 | 0.260929 | 3.458533 |
| Active power | 0.07304 | 2.32E-07 | 0.005005 | 0.41349 | 3.221059 |
| Opening degree of regulating door 1 | 0.237174 | 2.77E-07 | 0.010442 | 0.306821 | 3.572224 |

| Column name | Maximum error | Minimum error | Mean error | Missing ratio | Time used (s) |
|-------------|---------------|---------------|------------|---------------|---------------|
| Main control command | 0.085134 | 2.34E-05 | 0.017353 | 0.352163 | 10.13966 |
| Main steam pressure | 0.316095 | 1.29E-05 | 0.028357 | 0.260929 | 6.073386 |
| Active power | 0.401103 | 4.44E-06 | 0.03545 | 0.41349 | 5.652274 |
| Opening degree of regulating door 1 | 0.696386 | 8.74E-07 | 0.023393 | 0.306821 | 3.073117 |

According to the statistical results in Table 2 and Table 3, based on the combination of linear interpolation and LightGBM, it has obvious advantages in filling error and efficiency when performing data missing filling. The filling effects of the two methods are shown below. For the convenience of the display, only the first 20,000 points of each measuring point are used for display.
Table 4 Visualization of test results

| Method                  | Main control command | Main steam pressure | Active power |
|-------------------------|----------------------|---------------------|-------------|
| Random forest           |                      |                     |             |
| Linear interpolation    |                      |                     |             |
| LightGBM                |                      |                     |             |

---

**Random forest**

- Main control command
- Main steam pressure
- Active power

**Linear interpolation and LightGBM**

- Main control command
- Main steam pressure
- Active power
5. Conclusions

Through experimental tests, the data missing filling method based on linear interpolation and LightGBM algorithm indicate good performance in fitting error control and calculation efficiency. When performing LightGBM completion, this method first performs linear interpolation preprocessing, which greatly reduces the impact of missing independent variables on the filling error of dependent variables, and greatly improves the data utilization and complete accuracy of the model; secondly, when filling data, according to the missing ratio from high to low, the measuring points with the largest missing ratio will be filled first to further reduce the error caused by hard interpolation. In summary, the data filling method proposed in this paper can ensure a better filling effect when filling source network data and has a certain value for improving the accuracy of data mining.

References

[1] SHI Yanhua, LI Shuyu. Cleaning research of clustering feedback learning[J]. Computer Engineering and Applications, 2011, 47(30): 127-131.
[2] Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society, 1977, 39(1):1-38.
[3] Rubin D. Mutiple imputations for nonresponse in surveys[J]. New York: John Wiley & Sons, Inc, 1987.
[4] Schafer J, Olsen M. Multiple imputation for multivariate missing-data problems: a data analysis’s perspective[J]. Multivariate Behavioral Research, 1998, 33(4): 545-571.
[5] Chen S. Imputation of missing values using quantile regression. Dissertations & Theses – Gradworks, 2014.
[6] Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of Random Forest and Parametric Imputation Models for imputing Missing Data Using MICE: A CALIBER Study[J]. American Journal of Epidemiology, 2014, 179(6): 764.
[7] Masseran N, Razali A M, Ibrahim K, Zaharimand, Sopian K. Application of the single imputation method to estimate missing wind speed data in Malaysia[J]. Research Journal of Applied Sciences Engineering & Technology, 2013, 10:1780-1784.
[8] Madhu G. A Non-Parametric Discretization Based Imputation Algorithm for Continuous Attributes with Missing Data Values[J]. International Journal of Information Processing, 2014.
[9] GUOLK, QIM, FINLEY T, etal. Lightgbm: A highly efficient gradient boosting decision tree[C]// NIPS. Advances In Neural Information Processing Systems. USA:[s.n.], 2017: 3146-3154.