Missing Data estimation with a bi-dimensional adaptive weighted method for power grid data

Xi Wang¹, Mingwei Li², Jianjia Zhou³, Yan Deng⁴ and Qingqing Zhang⁵, *

¹,²,³,⁴ Sichuan Electric Power Company, State Grid Corporation of China, China
⁵ School of Automation Engineering, University of Electronic Science and Technology of China, China

*Corresponding author e-mail: 1556877687@qq.com

Abstract. The power grid data often contain missing values, which will seriously affect the analysis of the power grid development trend. To estimate missing values, the k-nearest neighbor and linear regression imputation method have been proved to be effective. These methods only consider the correlation between the samples or just consider the correlation between the features. However, we do not know which correlation of the two dimensions is more important in some special missing values. So this paper proposes a bi-dimensional adaptive weighted imputation method which considers the correlation of the two directions simultaneously. The experimental results show that the bi-dimensional adaptive weighted imputation method can reduce the estimation error.

1. Introduction

With the rapid development of the power industry, the large data analysis technology has become an increasingly powerful analysis tool to predict the development trend of power grid. Accurately predicting the development trend of the power grid can provide a strong scientific basis for the management and decision-making. So, it is important to ensure the integrity and accuracy of the data. But in real life, the data of the power grid is not all integrated. There will be a lot of missing data, because of human factors, such as errors in data entry, job loss or intentional forgery cause data loss; or because of objective factors, such as equipment failure and route interruption cause data loss. The missing data causes large amounts of information will be lost, seriously affect the data mining or statistical analysis. Therefore, the result of data analysis is likely to be very biased. In order to improve the accuracy of power grid data and ensure the integrity of data, it is necessary to solve the problem of missing data in power grid.

There are many ways to deal with missing data. The simplest way to deal with the missing data is ignore or simply replace missing values by zeros [1]. However, this method can cause data to lose a lot of information, resulting in poor performance. The mean imputation method is to fill the missing value by row average [2], but this method will artificially increase the noise of the data. There are lots of complex and good performance methods, such as regression model imputation [3], the Expectation Maximization (EM) imputation [4], neural network imputation [5], k-nearest neighbor (KNN) imputation [6]. KNN imputation method is one of the most popular and fastest methods to estimation missing value. But the KNN imputation method just considers the correlation between samples, because it considers the distribution space of samples. The linear regression imputation method utilitze
other complete data to estimate missing value. However when the variables are not linearly correlated or the prediction variables are highly correlated, they may lead to biased estimates.

In this paper we consider the correlation between the samples and the features by using the KNN imputation method and linear regression imputation method simultaneously. According to the correlation between the samples and features, we propose a bi-dimensional adaptive weighted imputation method. The missing values are estimated by the adaptive weights of the first nearest neighbor and linear regression value.

2. Proposed method

The method proposed in the paper utilizes the advantages of KNN imputation method and the linear regression imputation method to estimate missing value. This method constitute of three steps: firstly the nearest neighbor of missing value is estimated by the KNN imputation method; secondly the regression value of missing value is estimated by the linear regression imputation method; finally the missing value is imputed by the adaptive weights of the first nearest neighbor and linear regression value.

2.1. KNN imputation method

The KNN imputation method uses Euclidean distance to calculate the first nearest neighbor. The steps of the KNN imputation method to estimate the missing values are as follows:

Step1) firstly, the dataset \( D \) is shown in the Table1 (0 replace missing value). \( S_1, S_2, \cdots, S_m \) represent \( m \) samples and \( F_1, F_2, \cdots, F_k \) represent \( k \) features in Table1. The dataset is divided into two parts: \( D_a \) represents a collection of samples containing missing values. And the collection of the remaining samples is \( D_b \).

Step2) then, compute the distance between the missing values of a sample vector and all sample vectors in \( D_b \) respectively: let \( S_i = [x_{i1}, x_{i2}, \cdots, x_{ik}] \) represents a sample vector with missing value \( x_{il} \) in \( D_a \), let \( S_j = [x_{j1}, x_{j2}, \cdots, x_{jk}] \) represents a sample vector with complete data in \( D_b \). And then use Euclidean distance to measure the distance between the two samples. The function is shown as follows:

\[
d_{ij} = \sqrt{\sum_{l=1}^{k} (x_{il} - x_{jl})^2}
\] (1)

Step3) compare the \( d_{ij} \) between \( S_i \) and every \( S_j \), select the first nearest neighbor of \( S_i \) in \( D_b \).

Step4) finally use the first nearest neighbor of \( S_i \) as the estimated value \( x_{ij} \) of missing value.

Step5) repeat step2 and step4 until find out the estimated value of all missing data.

| \( S_1 \) | \( S_2 \) | \( \cdots \) | \( S_m \) |
|---|---|---|---|
| \( F_1 \) | \( F_2 \) | \( \cdots \) | \( F_k \) |
| \( x_{11} \) | \( x_{12} \) | 0 | \( \cdots \) | \( x_{1k} \) |
| \( x_{21} \) | 0 | \( x_{23} \) | \( \cdots \) | \( x_{2k} \) |
| \( \cdots \) | \( \cdots \) | \( \cdots \) | \( \cdots \) | \( \cdots \) |
| \( x_{m1} \) | \( x_{m2} \) | \( x_{m3} \) | \( \cdots \) | \( x_{mk} \) |
2.2. Linear regression imputation method

The linear regression for estimating the missing value is expressed as:

\[ f(x_{il}) = a^T x \]  

(2)

Where \( f(x_{il}) \) is the missing value estimation of sample \( S_i \). Let \( x = (x_{i1}, x_{i2}, \ldots, x_{ih}, \ldots, x_{ik}) \), \( h \neq l \); \( a \) is estimated by least square method, the function is:

\[ a = (X^T X)^{-1} X^T y \]  

(3)

let \( y = (x_{jl}, x_{j2}, \ldots, x_{jl}, \ldots, x_{jm}) \), \( j \neq i \). \( X = \begin{pmatrix} x_{i1} & x_{i2} & \cdots & x_{ih} & \cdots & x_{ik} \\ x_{j1} & x_{j2} & \cdots & x_{jh} & \cdots & x_{jk} \\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mh} & \cdots & x_{mk} \end{pmatrix} \), \( X^T X \) is full-rank matrix.

2.3. Bi-dimensional adaptive weighted imputation method

The bi-dimensional adaptive weighted (BAW) imputation method estimates missing value by weighting the imputation values of KNN imputation and linear regression imputation. The adaptive weight function is:

\[ w_a = \frac{e^{-d_{ij}}}{e^{-d_{ij}} + e^{-\sqrt{f(x_i) - y_{ij}}}} \]  

(4)

\[ w_b = \frac{e^{\sqrt{f(x_i) - y_{ij}}}}{e^{-d_{ij}} + e^{-\sqrt{f(x_i) - y_{ij}}}} \]  

(5)

where \( w_a \) is the adaptive weighted of KNN imputation method, \( w_b \) is the adaptive weighted of linear regression imputation method, \( y_{ij} \) is the true value of \( x_{ij} \).

Therefore, the BAW imputation method is expressed as:

\[ x_{ab} = w_a x_{il} + w_b f(x_{il}) \]  

(6)

where \( x_{ab} \) is the imputation value of missing value.

3. Experiment

3.1. Data Set and Evaluation Method

In order to verify the practicability of the proposed method in this paper, the power grid data is used as experimental dataset. The dataset contains 19 urban samples and 6 features. To simulate the dataset contained missing values, there are 6 missing values are selected randomly, and the dataset is shown as Table2. The bold print is missing value. \( X1, X2, X3, X4, X5, X6 \) are features respectively. The city3, city6, city11, city12, city16 contain missing values respectively.
Table 2. The dataset with missing data.

| city  | X1     | X2     | X3     | X4     | X5     | X6     |
|-------|--------|--------|--------|--------|--------|--------|
| city3 | 673.34 | 360    | 145.1  | 186    | 152.7  | 91.0198|
| city6 | 780.4  | 329.3  | 98.7   | 93     | 77.6   | 43.15  |
| city11| **854.68** | **370.5**  | 121.05 | 168    | 92.8   | 52.88  |
| city12| 646.2  | 127    | **214** | 333    | 180.08 | 134.96 |
| city16| 918.05 | 256.5  | 204.6  | 279    | **270.04** | 203.03 |

The average absolute error (MAE), average absolute percentage error (MAPE) and mean square error (RMSE) are used to estimate the performance of the algorithm.

\[
 MAE = \frac{1}{m} \sum_{j=1}^{m} |\hat{y}_j - y_j| 
\]  

(7)

\[
 MAPE = \frac{1}{m} \sum_{j=1}^{m} \left| \frac{\hat{y}_j - y_j}{y_j} \right| 
\]  

(8)

\[
 RMSE = \sqrt{\frac{1}{m-1} \sum_{j=1}^{m} \left( \frac{\hat{y}_j - y_j}{y_j} \right)^2} 
\]  

(9)

where \( \hat{y}_j \) represents the predicted value of the missing value, \( y_j \) represents the true value.

3.2. Experimental Result

The proposed method is compared with KNN imputation method and linear regression imputation method. The imputation results of the three imputation methods are shown in Table 3, Table 4 and Table 5 respectively. The bold prints are the estimation values in the three tables. From these three tables, we can draw a conclusion that the estimation values of the proposed method are closer with the true values. For example the estimation value of city 6 is 43.06, and the true value is 43.15. They just differ 0.09 that is good performance. While the estimation value of other two methods for city 6 are 52.88 and 39.9 respectively.

Besides the above discussions, from Table 6, it can be found that the MAE of the 5 cities for the BAW imputation method is much smaller than other two methods. It shows that the proposed method performs significantly better than the KNN imputation method and linear regression imputation method. In addition, the MAPE of the 5 cities, 0.034 for the BAW imputation method, 0.320 for KNN imputation method, 0.194 for linear regression imputation method, show that the proposed method is more accurate and robust. The RMSE of the BAW imputation method is 5.199, which is smaller than others. So the results display that considering the correlation of between samples and features simultaneously makes the estimation values more effective and closer to the true values.

Table 3. The estimated value of missing values with BAW imputation method.

| city  | X1     | X2     | X3     | X4     | X5     | X6     |
|-------|--------|--------|--------|--------|--------|--------|
| city3 | 673.34 | 360    | 145.1  | 186    | 152.7  | 91.0198|
| city6 | 780.4  | 329.3  | 98.7   | 93     | 77.6   | 43.06  |
| city11| **840.5** | **354.07**  | 121.05 | 168    | 92.8   | 52.88  |
| city12| 646.2  | 127    | 200.49 | 333    | 180.08 | 134.96 |
| city16| 918.05 | 256.5  | 204.6  | 279    | **279.4** | 203.03 |
### Table 4. The estimated value of missing values with KNN imputation method.

|     | X1   | X2   | X3   | X4   | X5   | X6   |
|-----|------|------|------|------|------|------|
| city3 | 673.34 | 360  | 145.1 | 144  | 152.7 | 91.0198 |
| city6 | 780.4  | 329.3 | 98.7  | 93   | 77.6  | 52.88  |
| city11 | 901   | 412  | 121.05 | 168  | 92.8  | 52.88  |
| city12 | 646.2  | 127  | 204.14 | 333  | 180.08 | 134.96 |
| city16 | 918.05 | 256.5 | 204.6 | 279  | 135   | 203.03 |

### Table 5. The estimated value of missing values with linear regression imputation method.

|     | X1   | X2   | X3   | X4   | X5   | X6   |
|-----|------|------|------|------|------|------|
| city3 | 673.34 | 360  | 145.1 | 205.3 | 152.7 | 91.0198 |
| city6 | 780.4  | 329.3 | 98.7  | 93   | 77.6  | 39.9  |
| city11 | 765.79 | 205.36 | 121.05 | 168  | 92.8  | 52.88  |
| city12 | 646.2  | 127  | 145.1 | 333  | 180.08 | 134.96 |
| city16 | 918.05 | 256.5 | 204.6 | 279  | 314.57 | 203.03 |

### Table 6. The estimation result of BAW, KNN and linear regression.

| Method    | MAE  | MAPE | RMSE  |
|-----------|------|------|-------|
| KNN       | 62.248 | 0.320 | 35.849 |
| linear regression | 55.160 | 0.194 | 38.996 |
| BAW       | 8.768  | 0.034 | 5.199  |

### 4. Conclusion

In this paper, the BAW imputation method is proposed by considering the influence of the correlation between the samples and features. The proposed method is used to estimate the missing value of the power grid data. And the results of the experiment show that the estimated missing values for the BAW imputation method are more accurate than the KNN imputation method and linear regression imputation method. The proposed method in this paper has a better estimation precision and provides a new method for the processing of the missing data of the power grid. So it makes the maximum use of the existing data and improves the electricity. The efficiency of data mining and analysis will effectively promote the construction and development of smart grid in China.

### References

[1] Jing X Y, Qi F, Wu F. Missing data imputation based on low-rank recovery and semi-supervised regression for software effort estimation, International Conference on Software Engineering. IEEE, 2017:607-618.

[2] Sen S, Das M N, Chatterjee R. A Weighted kNN approach to estimate missing values, International Conference on Signal Processing and Integrated Networks. IEEE, 2016.

[3] Arteaga F, Ferrer A. Framework for regression based missing data imputation methods in online MSPC[J]. Journal of Chemometrics, 2005, 19(8):439-447.

[4] Li Z, Liao Y. Impact of missing data on parameter estimation of EM algorithm under exponential distribution, International Conference on Automatic Control and Artificial Intelligence. IET, 2013:278-281.

[5] Ravi V, Krishna M. A new online data imputation method based on general regression auto associative neural network[J]. Neurocomputing, 2014, 138(11):106-113.

[6] Cai Q, Wu Q, Dong H. The research of missing value estimation of gene sequence based on improved KNN, International Conference on Computer Science & Education. IEEE, 2009:1235-1238.