A Missing Power Data Filling Method Based on Improved Random Forest Algorithm*

Wei Deng[1], *Yixiu Guo*[2], *Jie Liu*[2], *Yong Li*[2*], *Dingguo Liu*[3] *and Liang Zhu*[3]

(1. State Grid Hunan Power Company Limited Research Institute, Changsha 410007, China;
2. College of Electrical and Information Engineering, Hunan University, Changsha 410082, China;
3. State Grid Hunan Electric Power Company Limited, Changsha 410004, China)

Abstract: Missing data filling is a key step in power big data preprocessing, which helps to improve the quality and the utilization of electric power data. Due to the limitations of the traditional methods of filling missing data, an improved random forest filling algorithm is proposed. As a result of the horizontal and vertical directions of the electric power data are based on the characteristics of time series. Therefore, the method of improved random forest filling missing data combines the methods of linear interpolation, matrix combination and matrix transposition to solve the problem of filling large amount of electric power missing data. The filling results show that the improved random forest filling algorithm is applicable to filling electric power data in various missing forms. What's more, the accuracy of the filling results is high and the stability of the model is strong, which is beneficial in improving the quality of electric power data.

Keywords: Big data cleaning, missing data filling, data preprocessing, random forest, data quality

1 Introduction

In recent years, with the development of smart meters and intelligent data collection system, it has been possible to provide sufficient electric power data to the smart grids research, such as fault diagnosis and load forecasting in power system. Data mining has greatly promoted the development and progress of smart grid. However, due to the problems such as the meter failure or the communication errors, electric power data is not complete. At the same time, in the UCI public dataset in the field of data mining, more than 40% of the databases contain missing data [1]. Therefore, researching the method of filling missing electric power data is not only a prerequisite for the smart grid research, but also meaningful for improving the quality of data.

Power big data has the characteristics of a large number, high dimension and time series. At the same time, there are many forms of missing electric power data, some are missing dispersedly and some are missing in succession. For example, due to the electric meter failure, there may be multiple days of electric power data not collected. Or due to signal transmission error, electric power data failed to upload. The traditional method is to delete missing data or use median, average and mode to fill the missing data. However, the traditional method is not suitable for the cases where the missing data is large, because it will destroy the continuity, integrity and accuracy of the electric power data. In addition, commonly used traditional filling missing data methods include single imputation and multiple imputation [2-3]. Although the accuracy of filling missing data can be improved to a certain degree, these methods have a large amount of computation, which are inefficient. Besides, based on the traditional methods, statistical analysis, machine learning and other methods are adopted to fill the missing data, which improve the filling accuracy and efficiency. However, these results are mainly for traditional data types. Power big data generally has high dimensionality and time series and there are many missing forms. Therefore, these methods are not fully applicable to the missing power data filling. In Ref. [4], the researchers have adopted clustering method to filling the missing data. By defining the degree of difference in the set of constraint tolerances, the method judges the degree of dissimilarity of the missing data from the perspective of the set, thus achieving the missing data filling. This method takes less time to calculate but the optimal parameters of the algorithm is difficult to determine, which affects the

* Corresponding Author, Email: yongli@hnu.edu.cn
* Supported by the State Grid Power Company of Hunan Province Science and Technology Project (No.5216A517000U).
Digital Object Identifier: 10.23919/CJEE.2019.000025

accuracy of the filling missing data. In Ref. [5], the researchers used the iterative test method of model fitting residuals to test the noise points and missing values in the electric power data. Then, correcting the abnormal data and filling the missing data in the iterative process, thereby improving the efficiency of electric power data analysis.

In this paper, an improved random forest filling missing power data algorithm is proposed. Firstly, the missing data is initially filled by linear interpolation. Then, the original missing data is combined with the initial filling data into a matrix. Finally, a random forest filling model is built to fill the missing electric power data. It shows the improved random forest can be applied to filling a large amount of electric power data missing. The filling results indicates that the method is better combined with the characteristics of electric power data, where the filling accuracy is high and the stability of the filling model is excellent.

2 Missing forms of electric power data

In the power grid, the electricity information collection system can collect three-phase voltage, current, active power, reactive power and other data at 96 points per day (one point every 15 minutes), which may cause large electric power data missing due to data transmission errors or meter failure. The representation of missing data includes not only the NULL value in the database, but also the special value used to indicate the missing value. For example, in the electric power collection system, -2 is used to indicate that the value does not exist. At the same time, the missing data in power system is complicated, where the data missing rate and the distribution position of missing data are different. As shown in Fig. 1, the 5 points three-phase current of April 2 were not collected due to the failure of the electric meter. In the case of severe data missing, data points may be missing for several days, which affects the process of power big data analysis, leading to erroneous results in data mining. Therefore, it is meaningful to study these missing values. The quality of big data in the grid can be improved by effectively filling the missing values.

Defining the missing data forms is shown in the Tab. 1. X_{ij} represents the data, I_i and V_j represent the attribute variable and the '?' represents the missing

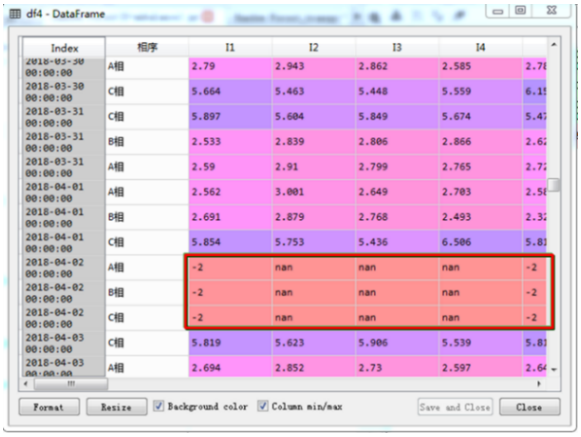


Fig. 1 Current data from the electricity collection system

data. The missing data forms include horizontal data missing, vertical data missing, discrete data missing, continuous data missing, single attribute data missing, multiple attribute data missing and all the attribute data missing. Single attribute data missing means that the data is only missing on one attribute V and the remaining attribute data is complete. Multiple attribute data missing is random, each attribute may have missing data and its distribution is irregular. All the attribute data missing means that all data on a column or a row is missing. Aiming at filling the missing power data, the paper proposes an improved random forest filling algorithm, which can be applied to fill all forms of missing data.

Tab. 1 The forms of missing data

Attribute variable	V_1	V_2	V_3	V_4	…	V_n
I_1	X_{11}	X_{12}	X_{13}	?	…	X_{1n}
I_2	X_{21}	?	X_{23}	?	…	X_{2n}
I_3	X_{31}	X_{32}	?	?	…	X_{3n}
I_4	?	?	X_{43}	?	…	X_{4n}
…	…	…	…	…	…	…
I_n	?	?	?	?	…	?

3 Improved random forest filling algorithm

3.1 Principle of random forest

In 2001, Breiman [6-7] combined Bagging theory with classification and regression decision tree and random subspace method to propose random forest algorithm. Random forest is an ensemble learning algorithm by combining multiple weak classifiers, the final result is voted or averaged. So the results will have higher accuracy and generalization

performance. The principle of filling random forest is shown in Fig. 2.

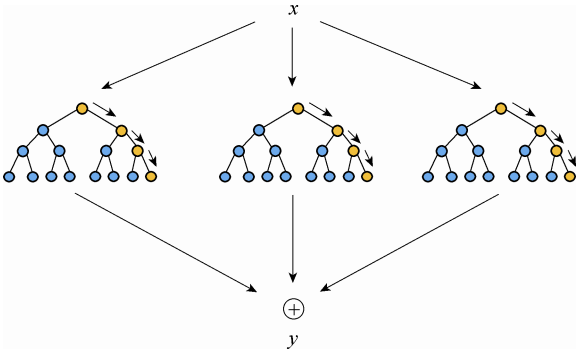


Fig. 2 The principle of filling random forest

Firstly, using Bootstrap resampling technique, multiple samples are randomly selected from the original training sample set x to generate a new training sample set. Then, multiple decision trees are constructed to form random forest. Finally, the random forest averages the output of each decision tree to determine the final filling result y. Due to the Bootstrap used in the random decision tree generation process, all samples are not used in a decision tree and the unused samples are called out of band [7-9]. Through out of band, the accuracy of the tree can be evaluated. The other trees are evaluated according to the principle and finally average the results, which is the principle of the random forest filling algorithm.

3.2 Algorithm of filling missing electric power data

In this paper, linear interpolation is used to fill the data for the first time, then the random forest algorithm is used to fill the missing electric power data. The specific algorithm process is as follows and the construction of data matrix are shown in Fig. 3.

(1) Preprocessing the data matrix $\boldsymbol{X}_{old}$. Put the column with the missing value at the forefront of the matrix, from the first column, the $\boldsymbol{X}_{sort}$ matrix is formed according to the number of missing values (small to large order).

(2) Because the random forest regression model cannot contain missing values, the linear interpolation of the interpolate module of SciPy is used to fill the missing data with the $\boldsymbol{X}_{sort}$ matrix to form the matrix $\boldsymbol{X}_{new}$.

(3) Splicing the first column of $\boldsymbol{X}_{sort}$ and $\boldsymbol{X}_{new}$ to remove all the columns except the first column to form a new $\boldsymbol{X}_{new}$ matrix. Only the first column of the matrix is a column with missing data and is also a target filling data column.

(4) Make a feature selection for the columns other than the target-filled column in $\boldsymbol{X}_{new}$. Iterating each feature by using out of band and the scores are sorted. The higher the score, the more important the feature. Then, the scores are sorted from high to low and the optimal feature data is selected by the accuracy of the model.

(5) Extract all the rows in the first column of the $\boldsymbol{X}_{new}$ matrix without missing data to form the Known matrix. At this moment, $\boldsymbol{X}$ is all columns starting from the second column in the Known matrix, $\boldsymbol{Y}$ is the first column in the Known matrix.

(6) Extract all the rows in the first column of the $\boldsymbol{X}_{new}$ matrix with missing data to form the Unknown matrix. At this moment, $\boldsymbol{Y}_{mis}$ is the first column with missing values in the Unknown matrix, $\boldsymbol{X}_{mis}$ is the columns remove the first column in the Unknown matrix.

(7) $\boldsymbol{X}$ is the feature attribute data, $\boldsymbol{Y}$ is the target (the target column to be filled) and the random forest regression model of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is established.

(8) Using the obtained model to perform unknown $\boldsymbol{Y}_{mis}$ prediction, which is using $\boldsymbol{X}_{mis}$ to predict $\boldsymbol{Y}_{mis}$.

(9) Update the $\boldsymbol{X}_{new}$ matrix with $\boldsymbol{Y}_{mis}$. and the first column is moved to the last column, moving the second column to the first column. Next, jump to step (3) for loop filling, as in the previous method until all columns with missing data are filled.

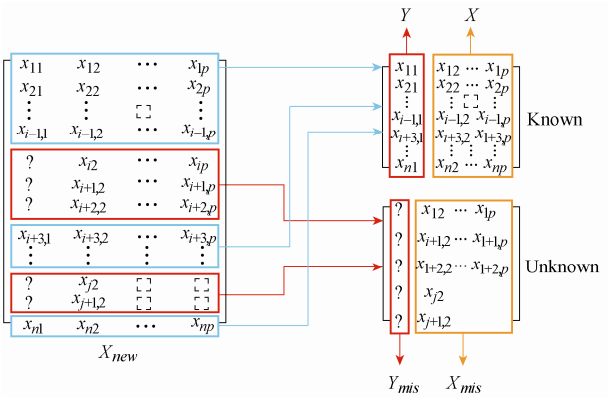


Fig. 3 The construction of data matrix

3.3 Filling missing data evaluation indicators

To analyse the performance of the filling missing electric power data, EVS (Explained variance score), MAE (Mean absolute error) and RMSE (Root mean

square error) are used as evaluation indicators [10-11], which are shown in the formulas (1) to (3).

$$EVS = 1 - \frac{Var(y_i - \tilde{y}_i)}{Var(y_i)} \tag{1}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \tilde{y}_i| \tag{2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2} \tag{3}$$

Where n is the number of test samples, y_i is the actual data, $\tilde{y}_i$ is the filling data, *Var* represents variance. These indicators can measure the accuracy of the missing data filling and the stability of the model.

4 Case study

In the paper, the data is derived from an actual distribution network in a certain area. The Python programming used to fill different missing forms of data through Lagrange interpolation, linear regression, spline interpolation and random forest.

4.1 Horizontal filling of discrete missing data

As shown in Fig. 4, the 96 points active power data on May 6 in 2018 is selected to be used to fill the missing data by deleting every other column. The 48 points in the middle are null, which is the discrete missing data. The matrix with horizontal discrete missing data is defined as $\boldsymbol{H}_1$. Then, the improved random forest is built to fill the deleted data.

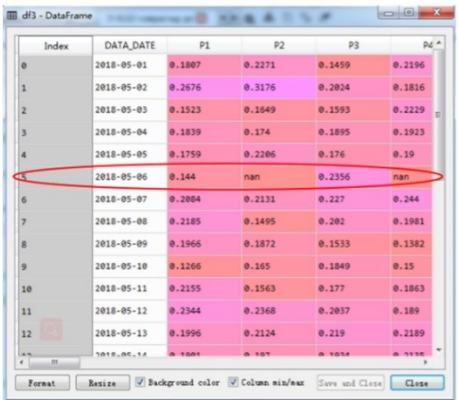


Fig. 4 The interlaced deleted active power data

The result of the filling is shown in Fig. 5. Comparing the actual data with the filled data, we can find that the filling result of the random forest fits the actual value best. It can be seen from Tab. 2 that the highest EVS of random forest is 0.892 and the error MAE and RMSE of random forest are the smallest. The EVS of Lagrange interpolation linear regression and spline interpolation are all negative. It shows that the random forest filling effect is the best.

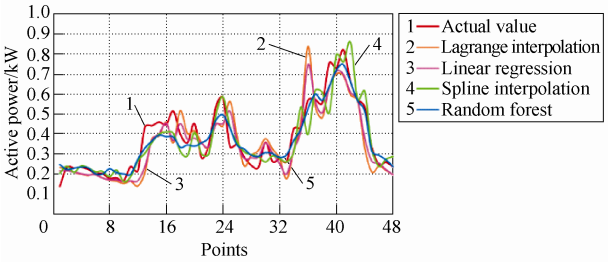


Fig. 5 The filling result and actual values

Tab. 2 The evaluation indicators of filling result

Evaluation indicators	Lagrange interpolation	Linear regression	Spline interpolation	Random forest
EVS	0.703	0.819	0.801	0.892
MAE	0.069	0.055	0.054	0.043
RMSE	0.094	0.075	0.075	0.056

4.2 Vertical filling of discrete missing data

As shown in Fig. 6, the P8 column is selected to be used to fill the missing data by deleting every other line. The deleted data has 20 points and the 20 points in the middle are null, which is the discrete missing data. The matrix with vertical discrete missing data is defined as $\boldsymbol{V}_1$. Then, the improved random forest is built to fill the deleted data.

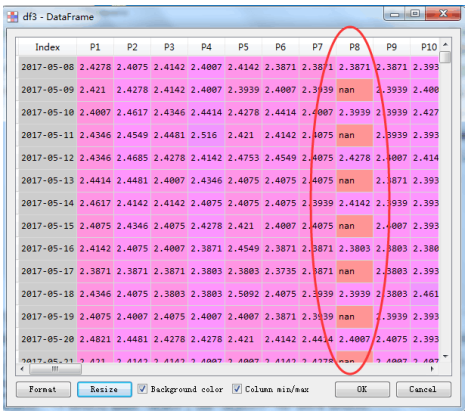


Fig. 6 The interlaced active power data

The result of the filling is shown in Fig. 7. Comparing the actual data with the filled data, we can find that the filling result of the random forest fits the actual value best. It can be seen from Tab. 3 that the highest EVS of random forest is 0.770 and the error MAE and RMSE of random forest are the smallest.

The EVS of Lagrange interpolation linear regression and spline interpolation are all negative. It shows that the random forest filling effect is the best.

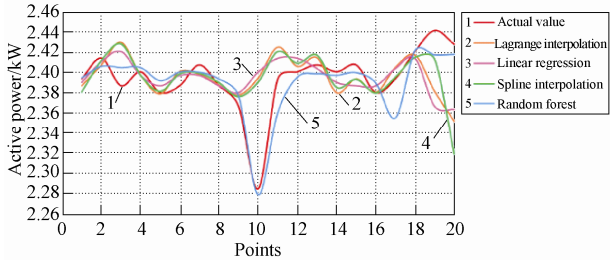


Fig. 7 The filling result and actual values

Tab. 3 The evaluation indicators of filling result

Evaluation indicators	Lagrange interpolation	Linear regression	Spline interpolation	Random forest
EVS	–0.391	–0.419	–0.489	0.770
MAE	0.022	0.022	0.021	0.011
RMSE	0.036	0.037	0.037	0.015

4.3 Horizontal filling of continuous missing data

In the electric power meter, there are many cases where data is missing for a certain day or for several days in a row, which is the continuous horizontal data missing. So we delete all the data for a certain day to do the filling research. The matrix with horizontal continuous missing data is defined as $\boldsymbol{H}_2$.

The results as shown in Fig. 8 and Tab. 4. Among the four filling methods, the curve of the random forest best fits the actual value. The EVS of the random forest is up to 0.649 and the error MSE and RMSE of random forest are the smallest. Relatively random forest filling has the best effect, followed by the linear regression method.

Tab. 4 The evaluation indicators of filling result

Evaluation indicators	Lagrange interpolation	Linear regression	Spline interpolation	Random forest
EVS	0.463	0.558	0.387	0.649
MAE	0.084	0.074	0.090	0.064
RMSE	0.115	0.103	0.123	0.091

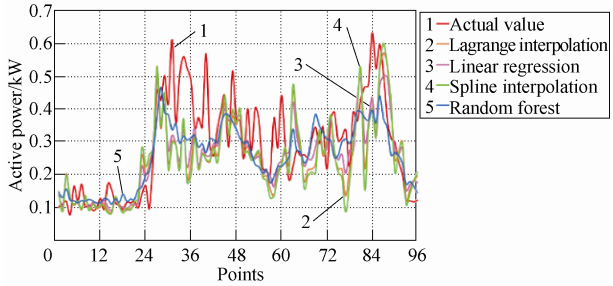


Fig. 8 The filling result and actual values

4.4 Vertical filling of continuous missing data

Generally, the filling effect of discrete missing data or horizontal continuous missing data is better. However, if the data is continuously missing vertically, the filling method may be invalid. The specific filling research is as follows: deleting 20 value of the vertical data consecutively as missing data for filling. The matrix with vertical continuous missing data is defined as $\boldsymbol{V}_2$.

The results as shown in Fig. 9 and Tab. 5. Lagrange interpolation, linear regression and spline interpolation methods are almost completely invalid, but the random forest can still fill the vertical missing data. It shows that the random forest filling model is more stable and has better generalization, which has excellent filling effect in various forms of data missing.

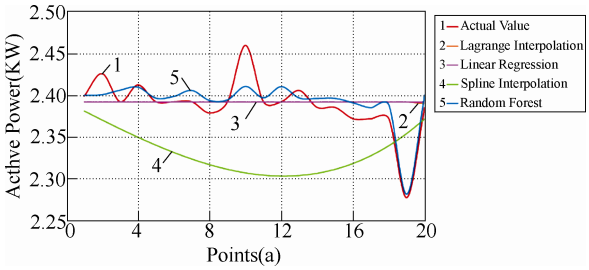


Fig. 9 The filling result and actual values

Tab. 5 The evaluation indicators of filling result

Evaluation indicators	Lagrange interpolation	Linear regression	Spline interpolation	Random forest
EVS	0.001	0.000	−0.796	0.755
MAE	0.018	0.018	0.065	0.013
RMSE	0.033	0.033	0.072	0.017

4.5 Filling for the transposing data matrix

Because the random forest is an algorithm for filling by columns, the program only needs to be iterated once to complete the vertical data filling. If the row of matrix has missing data, it needs to perform multiple loops to complete the data filling. The continuously missing data in the vertical column is selected to be filled. The vertical missing data matrix $\boldsymbol{V}_2$ are transposed becoming the horizontal missing data matrix $\boldsymbol{H}_2$ for filling.

It can be seen from Fig. 10 and Tab. 6, that though the number of iterations increases after

transposition, the error of filling data is smaller and the precision is higher. Since most of the power data is time series, transposing the vertical matrix to the horizontal matrix is more conducive to mining the time series of the data and improving the data filling effect. Therefore, the random forest algorithm for horizontally filling missing data is better.

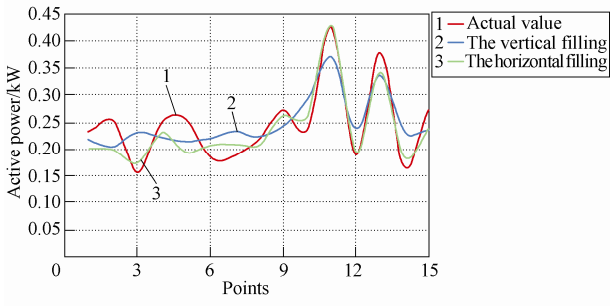


Fig. 10 The filling result and actual values

Tab. 6 The evaluation indicators of filling result

Evaluation indicators	Lagrange interpolation	Linear regression
EVS	0.606	0.852
MAE	0.041	0.024
RMSE	0.045	0.024
Iterations	1	15

5 Conclusions

In this paper, the Lagrange interpolation, linear regression, spline interpolation and random forest are used to fill the missing data of the active power data collected by the electric meter. The improved random forest algorithm works best, which can be applied to data filling in various missing forms with high precision and stable performance. At the same time, it is clear that the vertical missing matrix $\boldsymbol{V}_2$ can be transposed into the horizontal missing matrix $\boldsymbol{H}_2$, which can mining the time series of the data and improve the filling accuracy.

References

[1] Wei Ma, Yu Gu, Fangfang Li. Sequence-sensitive multi-source sensing data filling technology. *Journal of Software*, 2016: 2332-2347.

[2] T D Pigott. A review of methods for missing data. *Educ. Res. Eval.*, 2001, 7(4): 353-383.

[3] J C Wayman. Multiple imputation for missing data: What is it and how can I use it. Baltimore: Johns Hopkins University, 2003.

[4] Sen Wu, Xiaodong Feng, Zhiguang Dan. Missing data filling method based on incomplete data clustering. *Journal of Computers*, 2012, 35(8): 1726-1738.

[5] Yingjie Yan, Gewei Sheng, Shaopeng Qin. A big data cleaning method for power transmission and transformation equipment status based on time series analysis. *Automation of Electric Power Systems*, 2015, 39(7): 138-144.

[6] L Breiman. Bagging predictors. *Machine Learning*, 1996, 24(2): 123-140.

[7] L Breiman. Random forests. *Machine Learning*, 2001, 45(1): 5-32.

[8] T K HO. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, 20(8): 832-844.

[9] D H Wolpera, W G Macready. An efficient method to estimate bagging's generalization error. *Machine Learning*, 1996, 35(1): 41-45.

[10] R K Agrawal, F Muchahary, M M Tripathi. Long term load forecasting with hourly predictions based on long-short-term-memory networks. *IEEE Texas Power and Energy Conference*, 2018.

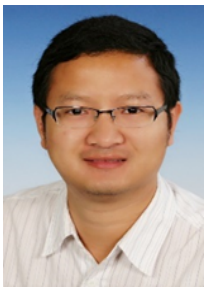
[11] Yinguo He, Shusheng Wu, Hong Zhou. Analysis of medium and long-term power demand forecasting in Hunan province. *Hunan Electric Power*, 2018, 38(6): 20-24.

Wei Deng was born in Hunan, China, in 1983. He received the Ph.D. degree in 2012, from the College of Electrical and Information Engineering, Hunan University, Changsha, China. Since 2012, he has worked in the State Grid Hunan Electric Power Co., Ltd. He has been engaged in the research of voltage and reactive power control, distribution network loss, and distribution network electric power research institute. His current research interests include optimal operation of smart distribution networks, application of big data and artificial intelligence in power grids, and distribution IOT technology.

Yixiu Guo was born in Fujian, China, in 1995. She received the B.Sc. degree in 2002, from the College of Electrical and Information Engineering, Hunan University, Changsha, China. She is currently a master student in College of Electrical and Information Engineering, Hunan University, Changsha, China. Her research interests include machine learning, smart grid and power system stability.

Jie Liu was born in Hunan, China, in 1979. She received the B.Sc. degree in 2002, from the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. She received the M.S. degree in 2006, from the School of Computer Science and Technology, Huazhong University of Science & Technology, Wuhan, China. She is currently a Ph.D. student in

College of Electrical and Information Engineering, Hunan University, Changsha, China. Her research interests include machine learning, modeling analysis and control of cyber physical power system, analysis and control of power quality.



Yong Li (S'09-M'12-SM'14) was born in Henan, China, in 1982. He received the B.Sc. and Ph.D. degrees in 2004 and 2011, respectively, from the College of Electrical and Information Engineering, Hunan University, Changsha, China. Since 2009, he worked as a Research Associate at the Institute of Energy Systems, Energy Efficiency, and Energy Economics (ie3), TU Dortmund University, Dortmund, Germany, where he received the second Ph.D. degree in June 2012. After then, he was a Research Fellow with The University of Queensland, Brisbane, Australia. Since 2014, he is a Full Professor of electrical engineering with Hunan University. His current research interests include power system stability analysis and control, ac/dc energy conversion systems and equipment, analysis and control of power quality, and HVDC and FACTS technologies.

Dingguo Liu was born in Hunan, China, on June 6,1979. He received the B.S. and M.S. degrees from the College of Electrical and Information Engineering, Hunan University, Changsha, China, in 2005 and 2008, respectively, where he has been working toward the doctoral degree in the College of Electrical and Information Engineering. His research interests include harmonics suppression, reactive power compensation, and smart grid.

Liang Zhu was born in Hunan, China, on August, 1974. He received the B.S. and M.S. degrees from the College of Electrical and Information engineering, Hunan University, Changsha, China. His research interests include power production management, distribution automation and power supply technology.