An Electricity Power Collection Data Oriented Missing Data Imputation Solution

Jiangqi Chen[1], Han Li[2,3(✉)], Ting Zhao[1], and He Liu[1]

[1] Advanced Computing and Big Data Technology Laboratory of SGCC, Global Energy Interconnection Research Institute, Beijing, China
{chenjiangqi, zhaoting, liuhe}@geiri.sgcc.com.cn
[2] College of Computer Science, North China University of Technology, Beijing, China
lihan@ncut.edu.cn
[3] Beijing Key Laboratory on Integration and Analysis of Large-Scale Stream Data, Beijing, China

Abstract. In Smart Grid, the incompleteness of electricity power collection data is gradually prominent. Thus, this paper presents an electricity power collection big data oriented missing data imputation solution, which comprises a big data processing framework and a missing data imputation method for power consumption data. Based on big data techniques, the given framework supports the large-scale electricity power collection data acquisition, storage and processing. To get a better result, the proposed method takes advantage of the correlation between the loss rate of power and the user power consumption. The feasibility and the effectiveness of the proposed method is evaluated. Experimental results show the proposed method is able to convert incomplete data set into complete data set, and has good imputation stability. Compared with the KNN algorithm, the proposed method has a lower imputation error, and is a positive attempt to combine domain-specific algorithms with traditional algorithms.

Keywords: Missing data imputation · Electricity power collection data
Big data · KNN · Data quality

1 Introduction

Being accompanied by the development of information techniques, communication techniques and IoT techniques, the total amount of data accumulated is rapidly increasing in nearly all fields. In order to make use of these data, more and more attentions have been paid to big data techniques [1] and data mining techniques in both academia and industry. Since data quality can have a significant effect on the conclusions that can be drawn from the data, it is necessary to preprocess the data [2].

Missing data occurs when no data value is stored for the variable in an observation [3]. Owing to the complexity of data accumulation, storage and processing, data missing is inevitable. Since data missing may reduce the data quality, it has become a challenging problem, and it is vital to replace missing data with substituted values.

I. Romdhani et al. (Eds.): CollaborateCom 2017, LNICST 252, pp. 243–252, 2018.
https://doi.org/10.1007/978-3-030-00916-8_23

Missing data imputation is a kind of technique that can convert incomplete data set into complete data set using appropriate strategies [4]. Recent years, a number of literatures show that the missing data imputation methods have begun to be discussed and applied in industry, medicine and other fields [4]. In the field of statistics, many mature missing data imputation methods have been proposed, such as the mean value method, the regression method, the hot-deck method, the expectation maximization method and so on. Although missing data imputation methods have been applied in a certain number of fields, it is still a challenging problem in the electric power industry. Thus this paper focuses on the missing data imputation method for the electricity power collection data.

Additionally, massive electricity power collection data is continuously produced by smart meters in the power grid of China. As an example, more than 400 million smart meters have been installed by the end of 2016. That is, the amount of electricity power collection data is rapidly increasing day by day. Therefore, a data processing framework for the electricity power collection data is required.

In conclusion, to solve the problem in missing data imputation for electricity power collection data, the missing data imputation method and the data processing framework are investigated in this paper.

2 Related Work

Data missing is one of the most ubiquitous and realistic problems in nearly every field. In the late 1970s, researchers begin to pay their attentions to resolve the problem of data missing. Recently, researchers at home and abroad have proposed plenty of missing data imputation techniques. Since most imputation methods are restricted to one type of variable whether categorical or continuous, a multi-objective genetic algorithm for missing data imputation MOGAImp is proposed in 2015 [5]. MOGAImp is proved to be effective and flexible, and is expected to be adapted in different application domains. In 2016, three missing data imputation methods based on fuzzy-rough methods are given [6]. Experimental results show these methods are effective. However, the time complexity is a little high and could be reduced. In 2017, a multiple imputation method in the presence of nonignorable nonresponse is proposed [7]. The multiple imputation method is verified to be relatively robust in bias and coverage. However, its efficiency is required to be improved. In addition to the traditional missing data imputation methods, missing data imputation methods which make use of domain knowledge has begun to draw more attention. In 2015, a hybrid prediction model with missing value imputation for medical data is presented [8]. The proposed hybrid model is the first one to use combination of K-means clustering with multilayer perceptron. Experimental results show the proposed method is feasible, robust and efficient. In 2016, a data imputation method based on the manufacturing characteristics is proposed for resolving the data missing problem in steel industry [9]. In the method, a correlation analysis method called NGCC and its corresponding model are used. The experimental results indicate that the proposed method is feasible and exhibits a better performance on the accuracy [9]. In 2017, a missing data imputation method based on genetic algorithms is given for questionnaires. In the method, Bayesian and Akaike's

information criterions are taken as the fitness functions [10]. Experimental results show that the proposed method is more effective than Multivariate Imputation by Chained Equations (MICE) algorithm [11]. As illustrated above, traditional missing data imputation methods are not universal, and making use of the domain knowledge is becoming a trend of the imputation of missing data. Thus, an electricity power collection data oriented missing data imputation method is proposed by combining the domain knowledge and the traditional missing data imputation method in this paper.

Furthermore, since the scale of electric power data rapidly increases, how to efficiently processing the electric power data has become a core issue in the construction of smart grid. In 2015, a Spark based unified cluster computing platform which is suitable for storing and performing big data analytics on smart grid data is proposed [12], and the feasibility of the platform is depicted. In 2017, a big data framework for analytics in smart grids is presented, which contains various big data techniques including Flume, HDFS, MapReduce, Hive, Tableau and so on [13]. By applying the framework on two scenarios to visualize the energy, the feasibility of the framework is verified. Since the electricity power collection data is a part of the electric power data, the scale of electricity power collection data is also large. Thus, this paper designs a big data processing framework to support the processing of electricity power collection data in the power grid in China.

3 The Proposed Big Data Processing Framework

In order to establish a proper big data processing framework for electricity power collection data, the characteristics of the power grid of China are summarized. (1) Large scale: In China, the size of the electric power grid is very large and is expanding year by year. Thus, the amount of the electricity power collection data rises rapidly. (2) Multilayer: There are multiple layers in the power grid of China, such as the headquarters, the provincial power grid and the prefectural power grid. From the perspective of the headquarters, the headquarters and the provincial power grid are two major layers. (3) Multiple data types: The electricity power collection data is divided into three main types, including the real-time data, the archive data and the historical data. The real-time data is continuously generated by smart meters. The archive data contains the information of users. The historical data is similar to the real-time data, but is generated during a period of time in the past. (4) Different performance requirements: According to the requirements of different electric power applications, the time to achieve an data processing varies from minute to hour.

As shown in Fig. 1, a big data processing framework for electricity power collection data is designed to accords with the above characteristics. In brief, the proposed framework is divided into two layers, can support large-scale data processing with MapReduce and Spark, and is able to accumulate and store the real-time data, the archive data and the historical data.

In the framework, the data acquisition subsystem focuses on accumulating data from intelligent meters, and transferring these data into data storage subsystem. In the data collection module, data are obtained by establishing long connections between the data collection module and the intelligent meters through socket. In the real-time data

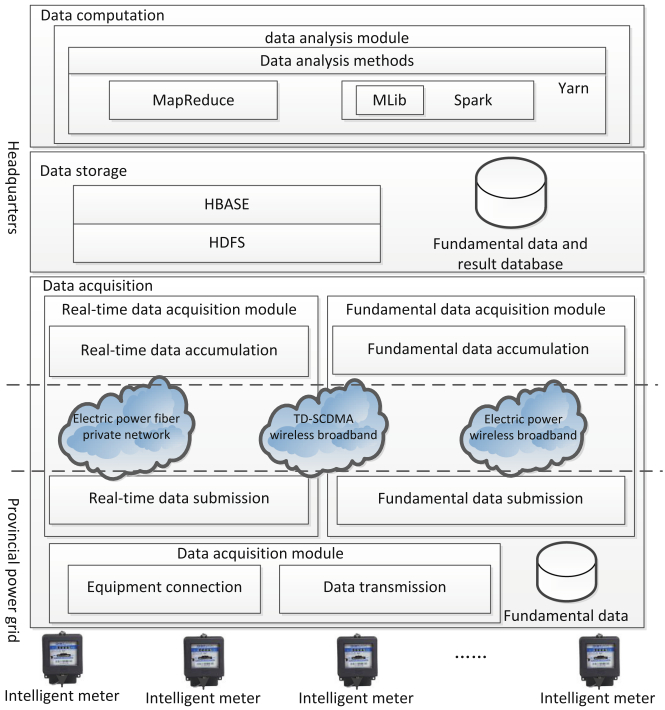

Fig. 1. The big data processing framework for electricity power collection data is divided into two layers including the provincial power gird and the headquarters, and is composed of three subsystems including the data storage subsystem, the data computation subsystem and the data acquisition subsystem.

acquisition module, Kafka is used for building real-time data pipelines. In order to transfer the fundamental data which refers to the history data and the archive data, an ftp service is applied in the fundamental data acquisition module. The data storage subsystem is an integrated storage environment, in which MySQL is used to save archive data, HDFS is applied to receive original electricity power collection data, and HBASE is used to save the parsed electricity power collection data. The data computation subsystem concentrates on the processing of electricity power collection data. Missing data imputation methods and a distributed computation environment based on MapReduce and Spark are the key components.

4 The Proposed Missing Data Imputation Method

In the electric power industry, power which is defined as the rate at which electrical energy is transferred by an electric circuit is a key indicator for user behavior analysis. Therefore, this paper takes the power consumption as the research object, and the data missing problem in this paper is defined as how to calculating the missing power consumption data.

4.1 Description of the Data Missing Problem

The description of the data missing problem in this paper is depicted as follows.

There are N users in a transformer area T. Every day, the smart meters record the power consumption of the transformer area for each user. Assuming the collection of power consumption occurs M times/day, $M \times D$ records are collected within D days. In this paper, (d, m) represents the mth time point in the dth day, P_{dmT} represents the output power of the transformer area T, P_{dm0} depicts the loss of power, the power consumptions of N users are respectively depicted by $P_{dm1}, P_{dm2}, \ldots, P_{dmN}$. Then P_{dm0} is set to $P_{dmT} - (P_{dm1} + P_{dm2} + \cdots + P_{dmN})$. As a result, the problem to be resolved in this paper is described as calculating the missing power consumption data for the ith user P_{dmi}.

4.2 Roadmap of the Proposed Method

In practical applications, any data entry whose power consumption is missing will be discarded. Due to the high missing rate of power consumption, the amount of data will be obviously reduced and the accuracy of further analysis will also decreased. Thus, it is necessary to calculate the missing power consumption. Since considering the domain knowledge during the imputation of missing data can get better result, a missing power consumption data imputation method is proposed by combing the KNN algorithm and the correlation between the loss rate of power and the user power consumption. In the proposed method, the KNN algorithm is used to construct a candidate set, while the domain knowledge is used to select a right candidate.

The technology roadmap of the proposed method is divided into three stages. In the first stage, the loss rate of power is calculated by estimating the state of transformer area T. In the second stage, the missing user power consumption is estimated according to the daily power curve. In the last stage, the result is adjusted based on the correlation between the loss rate of power and the user power consumption.

4.3 Estimation of the Loss Rate of Power

For the mth time point in the dth day, the ratio of the power consumption for user n to the total power consumption is represented by $r_{dmn} = P_{dmn}/P_{dmT}$, and the loss rate of power at a certain time is able to be estimated according to the ratio of the power consumption for each user to the total power consumption within the transformer area. The details are described as follows:

At time point (d, m), $[r_{dm1}, r_{dm2}, \ldots, r_{dmN}]$ is considered as a sample and is represented by R_{dm}. G is a set of time points with complete user power consumption. For $(d, m) \in G$, if all users' power consumption within the transformer area are known, R_{dm} is complete, the loss rate of power r_{dm0} can be calculated using formula (1), and $\{(R_{ij}, r_{ij0}) | (i, j) \in G\}$ is used to represents a known sample set.

$$r_{dm0} = P_{dm0}/P_{dmT} = 1 - (r_{dm1} + r_{dm2} + \cdots + r_{dmN}) \tag{1}$$

For each time point $(d, m) \notin G$, R_{dm} is incomplete, and the loss rate of power is unknown. So the loss rate of power for R_{dm} is estimated as follows:

There are u users which are represented as n_1, n_2, …, n_u. The time point is (d, m). R^*_{dm} which does not have any missing data is excerpted from R_{dm}, and is represented as $R^*_{dm} = [r_{dmn_1}, r_{dmn_2}, \ldots, r_{dmn_u}]$. R^*_{ij} is excerpted from R_{ij}, and can be depicted as $R^*_{ij} = [r_{ijn_1}, r_{ijn_2}, \ldots, r_{ijn_u}], (i,j) \in G$. Based on the KNN algorithm in which the city block distance is used, k_1 candidates $\{ R^*_{i_u j_u} \}_{u=1}^{k_1}$ are selected from $\left\{ R^*_{ij} | (i,j) \in G \right\}$, and the loss rate of power for R_{dm} is preliminarily estimated according to formula (2).

$$\widetilde{r_{dm0}} = \frac{1}{k_1} \sum\nolimits_{u=1}^{k1} r_{i_u j_u 0} \tag{2}$$

4.4 Estimation of the Missing Power Consumption

Based on the known daily power curves, the missing data of power consumption in an incomplete daily curve is able to be estimated as follows:

The daily power curve of user n in the dth day is represented by $L_{dn} = [P_{d1n}, P_{d2n}, \ldots, P_{dMn}]$. Due to the missing of P_{dmq}, L_{dq} is incomplete. Therefore, L^*_{dq} which does not have any missing data is excerpted from L_{dq}, and is represented as $L^*_{dn} = [P_{dm_1 q}, P_{dm_2 q}, \ldots, P_{dm_l q}]$. Assuming H_{ij} is the aggregation of daily power curves where both $\{P_{im_u j}\}_{u=1}^{l}$ and P_{jmi} do not have any missing data, $L^*_{ij} = [P_{im_1 j}, P_{im_2 j}, \ldots, P_{im_l j}], (i,j) \in H_{ij}$ which does not have any missing data is excerpted from H_{ij}, and $\{(L_{ij}, P_{imj}) | (i, j) \in H_{ij}\}$ is used to represents a known sample set. Based on the KNN algorithm in which the correlation distance is used, k_2 most candidates $\{ L^*_{i_u j_u} \}_{u=1}^{k_1}$ are selected, and the missing power consumption is estimated according to formula (3).

$$P^{(u)}_{dmq} = P_{i_u m j_u} \times \left(\sum\nolimits_{w=1}^{l} P_{dm_0 q}\right) / \left(\sum\nolimits_{w=1}^{l} P_{i_u m_w j_u}\right), u = 1, 2, \ldots k_2 \tag{3}$$

4.5 Adjustment of the Missing Power Consumption

Assuming the power consumptions of v users are not complete, the sum of missing power consumption can be calculated on the basis of the known user power consumption and the loss rate of power according to formula (4).

$$\sum\nolimits_{j=1}^{v} \widetilde{P_{dmq_j}} = (1 - \widetilde{r_{dm0}}) P_{dmT} - \sum\nolimits_{i=1}^{u} P_{dmn_i} \tag{4}$$

Based on the correlation between the power consumption and the loss rate of power, the final power consumption will be adjusted as follows:

At the time point (d, m) $\notin$ G, there are k_2 candidates $P^{(u)}_{dmq_j}, j = 1, 2, \ldots, v, u = 1, 2, \ldots, k_2$ for each absent user power consumption P_{dmq_j}. That is, there are k_2^v

candidates for v user power consumption. For each candidate, the difference between $\sum_{j=1}^{v} P_{dmq_j}^{u_j}$ and $\sum_{j=1}^{v} \widetilde{P_{dmq_j}}$ is calculated, and $\{ P_{dmq_j}^{0} \}_{j=1}^{v}$ is used to represent the candidate that have the minimum difference.

Then, the corresponding loss rate of power is calculated according to formula (5).

$$r_{dm0}^{0}=1-\left(\sum_{i=1}^{u} P_{dmn_i}+\sum_{j=1}^{v} P_{dmq_j}^{0}\right)/P_{dmT} \tag{5}$$

If $r_{dm0}^{0} \in [max(r_1, \widetilde{r_{dm0}}-r_2), \widetilde{r_{dm0}}+r_2]$, the final candidate of the loss rate of power $\widehat{r_{dm0}}$ is set to $\widetilde{r_{dm0}}$. If $r_{dm0}^{0} < max(r_1, \widetilde{r_{dm0}}-r_2)$, $\widehat{r_{dm0}}$ is set to $max(r_1, \widetilde{r_{dm0}}-r_2)$. If $r_{dm0}^{0} > \widetilde{r_{dm0}}+r_2$, $\widehat{r_{dm0}}$ is set to $\widetilde{r_{dm0}}+r_2$.

Next, based on $\widehat{r_{dm0}}$, the sum of the missing user power consumptions is calculated according to formula (6).

$$\sum_{j=1}^{v} \widehat{P_{dmq_j}} = (1-\widehat{r_{dm0}})P_{dmT}-\sum_{i=1}^{u} P_{dmn_i} \tag{6}$$

Finally, the missing data is recalculated according to formula (7).

$$\widehat{P_{dmq_j}} = [(1-\widehat{r_{dm0}})P_{dmT}-\sum_{i=1}^{u} P_{dmn_i}] \times P_{dmq_j}^{0}/(\sum_{h=1}^{v} P_{dmq_h}^{0}) \tag{7}$$

5 Experiments and Results

The experiments are designed to evaluate the feasibility and effectiveness of the proposed missing data imputation method. In the experiments, a realistic data set obtained from the smart meters in a certain transformer area of China is applied. For the purpose of calculating the imputation error, a subset of the data set, which has no missing data, is considered as the fundamental data set. Based on the three-phase characteristics of electricity power collection data, the fundamental data set is divided into three data sets, including Tran1_PhaseA, Tran1_PhaseB and Tran1_PhaseC. PhaseA, PhaseB and PhaseC respectively refers to the three phases of three-phase electric power. The scales of these data sets are the same, and they have the same attributes. Taking data set Tran1_PhaseA as an example, the basic information of Tran1_PhaseA is given in Table 1.

Table 1. Basic information of Tran1_PhaseA.

Number of attributes	Time slot (days)	Number of users	Frequency per day
10	50	44	48

To accomplish the experiments, the data sets should contain a certain number of missing data. The rate of missing data is defined as the percentage between the amount of missing data and the total amount of all data. In this paper, a data construction method is used to produce data sets with different rates of missing data.

Except for the rate of missing data, four other parameters are defined, which are the percentage of missing data for each time point (Thr_era2), the percentage of missing data for each daily power curve (Thr_era3), the percentage of time points without missing data (Thr_era4) and the percentage between the amount of complete daily power curves and the total amount of all power curves (Thr_era5). According to the practical experience, a fix value is set to each of the above parameters, which is listed in Table 2.

Table 2. Values of parameters.

Parameter	Thr_era2	Thr_era3	Thr_era4	Thr_era5
Value	80%	80%	5%	5%

The imputation error is defined as the mean value of the absolute error between the data set processed by the proposed method and the original data set. Assuming the data set which is processed by the proposed method and the original data set are respectively $ds1$ and $ds2$, the imputation error is calculated according to formula (8).

$$err\_rate = mean(mean(abs(ds1 - ds2))) \tag{8}$$

In general, the procedure of the experiments comprises three main steps. In the first step, the data construction method generates an incomplete data set by randomly deleting a certain percentage of data from the data set. In the second step, the proposed missing data imputation method is applied to the data set. In the last step, the imputation error is calculated.

To verify the feasibility of the algorithm, the proposed method is applied to data sets Tran1_PhaseA, Tran1_PhaseB and Tran1_PhaseC, and the rate of missing data of each data set is respectively set to 10%, 20%, 30%, 40% and 50%. Figure 2 shows the imputation errors of the experiments.

Based on KNN algorithm, the proposed missing data imputation method takes the loss rate of power into account when the missing user power consumption is calculated. In order to verify the effectiveness of the algorithm, the missing data are calculated

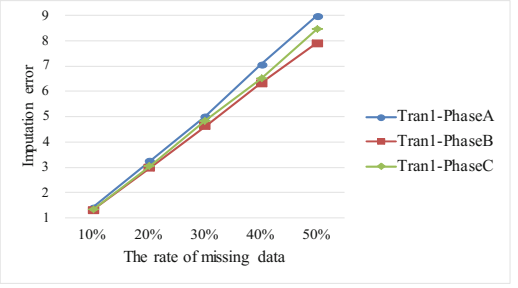

Fig. 2. The imputation errors of the experiments show that with the increasing of the rate of missing data, the imputation error tends to increase proportionally. That is, the proposed method is able to converts incomplete data to complete data, and is relatively stable.

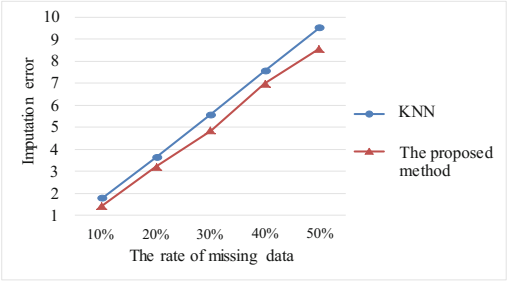

Fig. 3. The imputation errors of both the KNN algorithm and the proposed method shows that compared with the KNN algorithm, the proposed method has a lower imputation error. That is, it is able to get a data set with higher data quality.

using both the KNN algorithm and the proposed method. Figure 3 illustrates the imputation errors of both the KNN algorithm and the proposed method.

Due to the spatiotemporal correlation of the power consumption data,transformer area id, timestamp and user id are considered as the row key when storing data. Therefore, the proposed method is able to be supported by MapReduce or Spark. In addition, since the KNN algorithm is implemented on Mlib which is a machine leaning library provided by Spark, the efficiency of the proposed missing data imputation method is improved compared with that of a traditional serial program.

6 Conclusion

To solve the problem of missing data in the electric power industry, an electricity power collection big data oriented missing data imputation solution is investigated in this paper. The given solution consists of a big data processing framework for electricity power collection data and a missing data imputation method for power consumption data. The proposed framework accords with the characteristics of the power grid in China, and is able to support the acquisition, the storage and the computation of the electricity power collection data. Meanwhile, the proposed missing data imputation method calculate the missing values of user power consumption on the basis of KNN algorithm and the correlation between the loss rate of power and the user power consumption. Experimental results indicate the proposed missing data imputation method is a relatively stable method to converts incomplete data to complete data. Compared with the missing data imputation method which only bases on the KNN algorithm, the proposed method is able to generate more accurate values. Additionally, the proposed method has made a positive attempt to combine the traditional missing data imputation algorithm and the domain knowledge in the electric power industry. In the future, more professional knowledge is going to be applied to solve other data quality problems. The promotion and evaluation of the efficiency of the proposed solution is also the future work.

Acknowledgments. This paper is supported by research project of State Grid Corporation of China "Research on big data application technology and model in the company key areas" SGRIJSKJ(2016) 1104 project.

References

1. Ahmed, O., Fatima, Z.B., Ayoub, A.L., Samir, B.: Big data technologies: a survey. J. King Saud Univ. Comput. Inf. Sci. 1–18 (2017)
2. Ejaz, A., Ibrar, Y., Ibrahim, A., et al.: The role of big data analytics in Internet of Things. Comput. Netw. **129**, 1–13 (2017)
3. Song, Q., Shepperd, M.: A new imputation method for small software project data sets. J. Syst. Softw. **80**(1), 51–62 (2007)
4. Penny, K.I., Chesney, T.: Imputation methods to deal with missing values when data mining trauma injury data. In: The 28th International Conference on Information Technology Interfaces, pp. 213–218. IEEE Press, Cavtat (2006)
5. Fabio, L., Claudomiro, S., Igor, A., et al.: Multi-objective genetic algorithm for missing data imputation. Pattern Recognit. Lett. **68**, 126–131 (2015)
6. Mehran, A., Richard, J.: Missing data imputation using fuzzy-rough methods. Neurocomputing **205**, 152–164 (2016)
7. Im, J., Kim, S.: Multiple imputation for nonignorable missing data. J. Korean Stat. Soc. (2017, in press)
8. Archana, P., Sandeep, K.S.: Hybrid prediction model with missing value imputation for medical data. Expert Syst. Appl. **42**, 5621–5631 (2015)
9. Zheng, L., Jun, Z., Ying, L., Wei, W.: Data imputation for gas flow data in steel industry based on non-equal-length granules correlation coefficient. Inf. Sci. **367–368**, 311–323 (2016)
10. Celestino, O.G., Fernando, S.L., Francisco, J., et al.: Missing data imputation of questionnaires by means of genetic algorithms with different fitness functions. J. Comput. Appl. Math. **311**, 704–717 (2017)
11. Buuen, S., Groothuis, O.: MICE: multivariate imputation by chained equation in R. J. Stat. Softw. **45**, 1–67 (2011)
12. Shyam, R., Bharathi, G.H., Sachin, K.S., et al.: Apache spark a big data analytics platform for smart grid. Proc. Technol. **21**, 171–178 (2015)
13. Amr, A., Munshi, Y.A., Mohamed, R.I.: Big data framework for analytics in smart grid. Electr. Power Syst. Res. **151**, 369–380 (2017)