Electronics, Communications and Networks
A.J. Tallón-Ballesteros et al. (Eds.)

doi:10.3233/FAIA231171

Research on Imputation Method for Missing Values of Nuclear Power Plants' Operation Data

Jie LIU[a,1], Xiaodi QI[a], Zhiqiang WU[a,b], Wei JIANG[b], Zhi CHEN[a,b], Yifan JIAN[b] and Juan WEN[a]
[a] *School of Computer/Software, University of South China, China*
[b] *Laboratory of Nuclear Reactor System Design Technology, China Nuclear Power Research and Design Institute, China*

Abstract. The collection of nuclear power plants operating data is the basis for subsequent fault diagnosis and obtaining the operating status of nuclear power plants, but equipment failures and external interference will lead to missing operating monitoring data, which will reduce the quality of the data and thus reduce the accuracy of the subsequent analysis results. To solve this problem, this paper utilizes the fact that nuclear power plants have accumulated a large amount of operational data and researches the method of generating adversarial imputation network (GAIN)--based imputation method for missing values of nuclear power plants' operational data. The generator in the model estimates the missing values by learning the distribution of the true values, and the discriminator in the model discriminates which values are true and which are generated with the help of a hint matrix. The hint reveals partial information about the missing original samples to the discriminator, which the discriminator uses to focus its attention on the quality of the imputation of particular data values. Finally, a training set and a test set were constructed for comparative experiments on the PCTran simulation platform by simulating the operational data of the AP1000 as an example. The experimental results demonstrate that the investigated algorithm achieves lower root mean square error (RMSE), verifying the feasibility and accuracy of the method.

Keywords. Nuclear Power Plants, Missing Value Imputation, Deep Learning, Generative Adversarial Imputation Networks

1. Introduction

Numerous sensors are positioned throughout the nuclear power plant to track the state of various important pieces of equipment or systems as informationization of the facility continues to advance, taking safety, control, and other considerations into mind. The collection, transmission, and storage of operation data of nuclear power plants are completed mainly by automated instruments. The operator can assess the device's status and take prompt action to minimize human mistakes by studying the operating data.

However, because a nuclear power plant system's data gathering, transmission, and storage equipment operates in harsh environments, breakdowns, and outside interference

[1] Corresponding Author, Jie Liu, Associate Professor, School of Computer/Software, University of South China, China; jliuhn@foxmail.com.

can result in data quality issues like missing, drifting, and jumping in operation monitoring [1]. Data quality greatly affects the application and value of data, and poor data quality can lead to ineffective utilization of data [2]. The phenomenon of missing data is one of the more serious problems [3]. Therefore, it is necessary to investigate the method of imputation of missing values in the operational data of nuclear power plants.

The goal of missing value imputation of nuclear power plant operation data is to develop an effective and practical technique so that the imputation values for the operation parameter can accurately reflect the nuclear power plant's actual operational condition. In this paper, the Generative Adversarial Imputation Networks (GAIN) imputation algorithm for generating missing values of the operating data of nuclear power devices is studied, and the algorithms are verified by comparison. This is done based on an analysis of the characteristics of the operating data of nuclear power devices and the commonly used data missing value-filling algorithms.

2. Data characterization and selection of imputation methods

2.1. Characterization of operational data for nuclear power plants

The main pumps, steam generators, voltage stabilizers, and other equipment are all monitored by numerous sensors arranged in the reactor system of nuclear power plants. Each sensor measures the corresponding key parameters and returns and records them as time series [4]. Nuclear power plant operating parameters describe the nuclear power plant's operational state, and each nuclear power plant's operational state must precisely match a particular set of operating parameters of the data combination. Nuclear power plants currently contain a significant amount of operational data that can be automatically analyzed and used since they have the necessary data resources and conditions.

Missing data in a nuclear power plant refers to the occurrence when data gathered by the instrumentation and control system of a nuclear power plant is lost as a result of events like power outages or communication breakdowns, which causes the loss of data acquired by each sensor. Three categories of missing data exist [5,6]: Missing completely at random (MCAR), Missing at random (MAR), and Missing not at random (MNAR). According to MCAR, the values that are missing are entirely random and unrelated to the values that are present but not missing. MAR denotes that other observable values, rather than the missing values themselves, are related to the phenomena of missing data. MNAR denotes that both the missing and the observed variables are concurrently responsible for the missing values.

2.2. Selection of imputation methods

A large number of data imputation methods have been proposed by many researchers in the field of data imputation. The imputation methods for missing data can be categorized into three types [7]: (1) simple data-driven, (2) model-based, and (3) deep learning-based. Simple data-driven imputation methods include mean, median, mode imputation, and so on. Although mean, median, and mode imputation methods are simple and convenient, they may change the variance of the original data, and so on [8].

Regression-based [9] and K-nearest Neighbor (KNN)-based [10] approaches are two examples of model-based imputation techniques. The goal of regression-based imputation techniques like linear regression is to construct a regression model utilizing

relationships between variables; however, predictions are the same for samples with the same independent variables, so this might cause distortions in the sample distribution. KNN-based imputation methods select the K samples from the dataset that are most similar to the sample containing the missing value by calculating some similarity measures (e.g., Euclidean Distance, Mahalanobis Distance) and then use the values of these samples to estimate the missing value [10]. But for each missing value, it searches the entire dataset [11], which causes them to become very slow when dealing with large datasets.

Deep learning techniques have been widely used in the field of data imputation recently and have demonstrated significant promise [12]. Generative Adversarial Networks (GAN) [13] is a class of generative models that specialize in learning mappings from latent spaces to actual data distributions and are a better option for modeling data distributions. Human Gastrointestinal tract Abnormalities Network (HGANet) [14] is used to solve the problem of gastrointestinal anomaly identification. It did not require hand-crafted features, was trained end-to-end, and it learned directly the solution of a gastrointestinal abnormalities problem with endoscopic images.The conventional GAN architecture is used by GAIN [15], which functions effectively even with an incomplete dataset. GAIN adds a hint vector to help the discriminator verify that the generator generates the samples by the real underlying data distribution.

The algorithmic summary above shows that the GAN-based imputation methods can effectively use a large number of existing nuclear power plant operating parameters that do not have missing values for training and are effective at learning the distribution of the original data and Imputation the missing values. Because of this, the missing values for the operating parameters of nuclear power plants are Imputation in this paper using GAIN.

The GAIN imputation algorithm for missing values of operational parameters of nuclear power plants consists of the following steps:

(1) Normalization of nuclear power plants operational data to the range of 0 and 1;

(3) Construct the objective and loss functions of the model;

(2) For iterative training, a small sample of nuclear power plants operational data is chosen;

(4) The trained model imputes missing values in the operational data of the unit.

The brief process is shown in Figure 1.

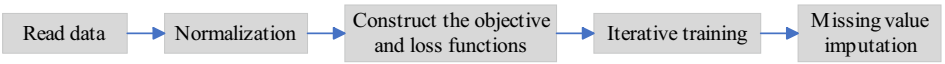

Figure 1. Sketch of the flow of the missing value imputation algorithm.

3. Algorithm design

3.1. Problem form of missing data imputation

Suppose the dataset is d-dimensional, which we denote as $X = (X_1, \ldots, X_d)$ and call it a data vector, and we denote its distribution as $P(X)$. M is the mask vector of the dataset, taking the value $\{0,1\}^d$ of the variable, denoted as $M = (M_1, \ldots, M_d)$. The relationship between the variables $\tilde{X}$, X, and M are as follows when we establish a new random variable $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_d)$:

$$\tilde{X}_i = \begin{cases} X_i, & M_i = 1 \\ *, & M_i = 0 \end{cases} \tag{1}$$

where $M_i = 1$ denotes observation of X_i, otherwise, it denotes absence of X_i. Therefore, M can be obtained using $\tilde{X}$.

3.2. Normalization of operational data for nuclear power plants

Essential parts of GAIN include the generator G and discriminator D, both of which are modeled as fully connected neural networks. For inputs in the [0,1] range, their activation functions, such as Sigmoid and ReLU, often perform better. The activation function may get saturated if the input data has a wide or irregular range of values, which can make the gradient disappear or burst and complicate training. Additionally, normalization ensures that the generated data falls within a reasonable range because it scales the G's output to the [0,1] range using the Sigmoid function.

The normalization formula for the data is shown below:

$$X_j' = X_j - min(X_j) \tag{2}$$

$$\tilde{X} = X_j'/(max(X_j') + 10^{-6}) \tag{3}$$

where X_j stands for each column of data in the dataset, $min(X_j)$ for each column's minimum value, X_j' for each column's values after subtracting the minimum value, $max(X_j')$ for each column's maximum value, 10^{-6} for preventing division by zero, and $\tilde{X}$ for the data after normalization.

3.3. Construction of model objective function and loss function

The extremely big very tiny value problem can be used to describe the GAIN method's training procedure. To increase the likelihood that M is accurately predicted, we first train the D. The G is then trained to reduce the likelihood that D will properly anticipate M. Thus, Eqs. (4) and (5) illustrate the model's goal function.

$$\min_G \max_D V(D,G) \tag{4}$$

$$V(D,G) = E_{\hat{X},M,H}\left[M^T \log D\left(\hat{X},H\right) + (1-M)^T \log\left(1 - D\left(\hat{X},H\right)\right)\right] \tag{5}$$

where the hint mechanism in GAIN, the random variable H, derives its value from the cueing space $\mathcal{H}$. D is then mathematically expressed as $D: \mathcal{X} \times \mathcal{H} \to [0,1]^d$ after receiving H as an extra input, where the i-th component of $D(\hat{x},h)$ corresponds to the probability of predicting the i-th component of $\hat{x}$ to be true given $\hat{X} = \hat{x}$ and $H = h$. H can be obtained using Eq. (6):

$$H = B \odot M + 0.5 \odot (1 - B) \tag{6}$$

where $B \in \{0,1\}^d$ is the random variable obtained by sampling k uniformly from $\{1,2,\ldots,d\}$ and applying Eq. (7). The term 0.5 in Eq. (6) represents a hint value similar to that used by Yoon [15].

Two components make up the loss function of the model: the loss function of the D and the loss function of the G. There are two sections to the G's loss function. First, because the output of G includes both estimates of missing data and estimates of non-missing values, the loss function of G is divided into two sections. Thus, the first part is the loss of missing values, while the second part is the loss of observations. The combined loss function $\mathcal{L}_G$ is given in Eq. (7).

$$\mathcal{L}_{\mathcal{G}} = \sum_{\forall i:b_i=0}(1-m_i)\log(\widehat{m}_i) + \alpha\sum_{j=1}^{d} m_i L_{obs}(x_i, x_i') \tag{7}$$

where α is a positive hyper-parameter, $b_i = 0$ corresponds to those values of $\widehat{M}$ for which H is 0.5 according to Eq. (6). $L_{obs}(x_i, x_i')$ is given in Eq. (8).

$$L_{obs}(x_i, x_i') = \begin{cases} (x_i - x_i')^2 & \text{if } x_i \text{ is continuous} \\ -x_i log(x_i') & \text{if } x_i \text{ is binary} \end{cases} \tag{8}$$

The output of the D can be expressed as $\widehat{M} = D(\hat{X}, H)$, thus the loss function of D can be represented by the cross-entropy Eq. (9).

$$\mathcal{L}_{\mathcal{D}} = \sum_{\forall i:b_i=0}[m_i\log(\widehat{m}_i) + (1-m_i)\log(1-\widehat{m}_i)] \tag{9}$$

3.4. Iterative training using small batches of data

We first use small batches of size 128 to optimize the D with a fixed G to maximize the probability that D correctly predicts M. The input to G consists of small batches of size 128, a noise matrix, and a corresponding mask matrix (0 for missing and 1 for non-missing), and G outputs the small batches that have been imputation. Second, we use the latest updated D to optimize the G to make the data generated by G more realistic, even if the probability that D correctly predicts M is minimized. The input of D consists of a hint matrix and a small batch of outputs from G, which outputs the probability that each value is true. G and D are trained iteratively, updating the parameters in G and D by backpropagation. Until the loss function converges or the training is complete, the trained model is obtained, and then the data with missing values are fed into the model to complete the imputation.

3.5. Missing data imputation process

The specific procedure of the method for imputation missing data of operating parameters of nuclear power plants is as follows, based on the prior design:

(1) Read the training dataset;

(2) Normalize the dataset;

(3) Construct the objective and loss functions of the model following the methodology in Section 3.3;

(4) Some data were arbitrarily and randomly eliminated using the operating parameters of the nuclear power plants that did not have missing values, and a small batch of size 128 was chosen to train the model;

(5) For each result produced using small batches of training, the losses of the G and the D are computed via Eqs. (7) and (9) for the G and the D, and then the parameters in G and D are updated by backpropagation;

(6) Imputation of nuclear power plants' operating parameters containing missing values using trained models.

The detailed flow of the missing value imputation technique for nuclear power plants' operating parameters is shown in Figure 2.

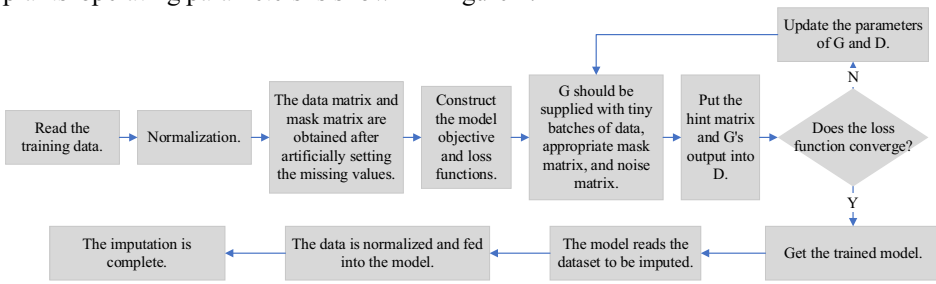

Figure 2. Data missing value imputation algorithm.

4. Experiments and results

We simulate and run the AP1000 run data using the PCTran simulation platform as a sample to experiment to confirm the accuracy and superiority of the suggested method. With constant values removed, all the operating parameters for the whole 2238-second period were chosen as the dataset, and 80% of them were randomly chosen for each run to serve as the training set and the remaining 20% as the test set. Six different missing rate thresholds—5%, 10%, 15%, 20%, 25%, 30%, and 35%—were established. The data imputation effect is assessed using the root mean square error (RMSE), which is defined as follows in Eq. (10):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{10}$$

where y_i denotes the true value, $\hat{y}_i$ denotes the imputation value, and n denotes the number of missing values.

Table 1. Comparison of the effect of imputation of missing values of the operating parameters of nuclear power plants. (*Optimal results are shown in bold.)

Missing rates (%)	5%	10%	15%	20%	25%	30%	35%
MM	0.2579	0.2572	0.2589	0.2586	0.2582	0.2588	0.2587
Hot Deck	0.0816	0.0822	0.0820	0.0819	0.0821	0.0816	0.0819
Soft-impute	0.0611	0.0614	0.0640	0.0648	0.0663	0.0702	0.0801
GAIN	**0.0547**	**0.0557**	**0.0565**	**0.0578**	**0.0624**	**0.0680**	**0.0769**

The Mean Value Imputation (MM), Hot Deck Imputation (Hot Deck), and Spectral Regularization Algorithms (Soft-impute) [16] are used as the control group to compare the Imputation impact of the three methods to reflect the superiority of the ones designed in this study. Each method is applied ten times at each missing rate, and the final value is determined by averaging the outcomes. The performance of our methods and the

comparator methods is shown in Table 1 over a range of data missing rates, from 5% to 35%. The strategies examined in this paper perform much better at all missing rate levels. Figure 3 demonstrates the trend of the imputation accuracy of the three methods as the missing rate increases.

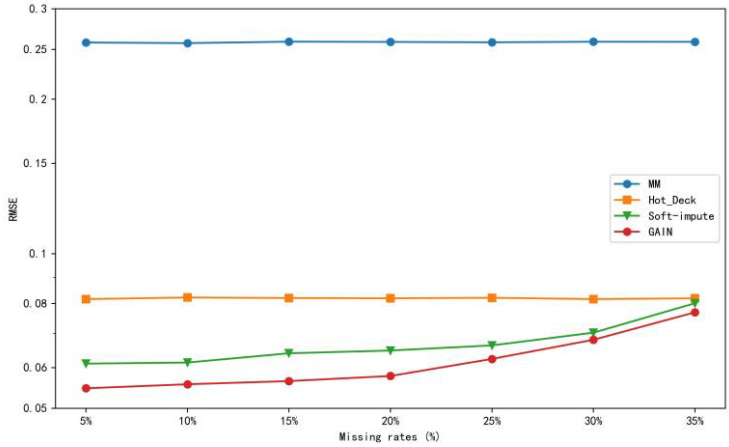

Figure 3. Performance of all methods at different deletion rates.

From Figure 3, it can be seen that the MM method and the Hot Deck method perform relatively smoothly with the increase of the data missing rate, while the method studied in this paper shows a slowly increasing trend in RMSE with the increase of the missing rate, i.e., the accuracy of the imputation gradually decreases. The underlying reason is analyzed because the MM method uses the average value of each sequence to imputation the missing values, and the average value of each sequence does not fluctuate drastically with the increase of the missing rate of data. The Hot Deck method fills in the missing values by randomly selecting the data of one sample from other samples with similar characteristics, and the values of the same nuclear power plant operating parameters have a high degree of similarity under the same conditions, so it is relatively easy for Hot Deck to find samples with similar characteristics to impute the missing values. However, it can be seen that the method studied in this paper still shows great advantages when the missing rate is relatively low.

5. Conclusion and analysis

The accuracy and stability of this method are demonstrated by comparing the experiments with the MM method and Hot Deck method, which show that the imputation accuracy of the method studied in this paper is significantly higher than that of the two compared methods when the data missing rate is low. The experimental findings demonstrate that the method investigated in this paper gradually loses data imputation accuracy as the data missing rate rises, however, our method still performs better below a missing data rate of 35%. To confirm the usefulness of the method examined in this paper, we will take into account the case where many missing modes occur at the same time in the upcoming research effort.

Some of the application prospects of the research methodology of this paper are as follows:

(1) The problem of missing values in offline data, mainly serves the function of data cleansing before data analysis and research, which is used to improve the quality of data and reduce the complexity of data analysis and research.

(2) For online applications, it can be applied in scenarios where data cannot be collected or transmitted when sensors are malfunctioning or failing, and the pre-stored data is utilized by this algorithm to train the model, which in turn complements the data.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62003157), the Open Subject of State Key Laboratory of Nuclear Reactor System Design: SQ-KFKT-24-2021-006, and the Scientific Research Fund of Hunan Provincial Department of Education (22C0223, 21B0434).

References

[1] Wang T, Yu R, Peng Q. Study on missing data filling algorithm of nuclear power plant operation parameters. Science and Technology of Nuclear Installations. 2022.
[2] Guo Z, Zhou A. Research on data quality and datacleaning: a survey. Journal of Software. 2002;13(11): 2076-2082.
[3] Deng JX, Shan LB, He DQ, Tang R. Processing method of missing data and its developing tendency. Stat. Decis. 2019;35:28-34.
[4] Zhang S, Lu T, Zeng H, Xu C, Zhang Z, Huang Q, Zhang X, Wang Y. Multi-feature fusion multi-step state prediction of nuclear power sensor based on LSTM. Nuclear Power Engineering. 2021;42(4):208-213.
[5] Kwak SK, Kim JH. Statistical data preparation: management of missing values and outliers. Korean Journal of Anesthesiology. 2017;70(4):407-411.
[6] Schlomer GL, Bauman S, Card NA. Best practices for missing data management in counseling psychology. Journal of Counseling psychology. 2010;57(1):1.
[7] Farhangfar A, Kurgan LA, Pedrycz W. A novel framework for imputation of missing values in databases. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans. 2007;37(5):692-709.
[8] Little RJA, Rubin DB. Statistical analysis with missing data. John Wiley & Sons; 2019.
[9] Ankaiah N, Ravi V. A novel soft computing hybrid for data imputation. In: Proceedings of the International Conference on Data Science. 2011, p. 1.
[10] Wang L, Fu D. Estimation of missing values using a weighted k-nearest neighbors algorithm. In: 2009 International Conference on Environmental Science and Information Application Technology. 2009;3: 660–663.
[11] Yoon J, Zame WR, van der Schaar M. Estimating missing data in temporal data streams using multi-directional recurrent neural networks. IEEE Transactions on Biomedical Engineering. 2018;66(5):1477-1490.
[12] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997;9(8):1735-1780.
[13] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. Advances in neural information processing systems. 2014;27.
[14] Iqbal I, Walayat K, Kakar MU, Ma J. Automated identification of human gastrointestinal tract abnormalities based on deep convolutional neural network with endoscopic images. Intelligent Systems with Applications. 2022;16:200149.
[15] Yoon J, Jordon J, Schaar M. Gain: Missing data imputation using generative adversarial nets. In: International Conference on Machine Learning. 2018. p. 5689–5698.
[16] Mazumder R, Hastie T, Tibshirani R. Spectral regularization algorithms for learning large incomplete matrices. The Journal of Machine Learning Research. 2010;11:2287-2322.