2022 21st International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Machine Learning enabled Missing Measurement Data Detection and Recovery of Electricity Grids

Min He
State Gird Ningbo Power Supply Company
Ningbo 315000, China
xyf8870@sina.com

Simeng Zheng
State Gird Ningbo Power Supply Company
Ningbo 315000, China
ning_prynne@163.com

Jia Yang
State Gird Ningbo Power Supply Company
Ningbo 315000, China
yangjia061@163.com

Yinghe Lin
Zhejiang Huayun Information Technology Co., Ltd.
Hangzhou, China
airlinyinghe@126.com

***Abstract*—This paper proposes a machine learning enabled missing data detection and recovery of electrical measurements based on the improved CPCAE. The proposed solution firstly accurately models the missing generation process to generate the missing mask and then combines the absolute difference sequence and the linear correlation as criteria to detect the possible missing segments under different signal-noise ratios (SNR). The solution divides the detected missing mask into different grades and reshapes the origin of one-dimensional data and mask into two-dimensional matrices as a kind of data enhancement. Then we intuitively turn to the deep learning technologies on image processing and design an improved CPCAE model to repair the damaged images. The proposed machine learning-enabled missing data detection and recovery solution are assessed through simulations and the numerical results confirmed its effectiveness for different missing situations.**

Keywords- data detection; data recovery; deep learning; cascaded pure convolutional auto encoder (CPCAE); two-dimensional reconstruction of one-dimensional data;

I. INTRODUCTION

Currently, along with the technological advances in information and communication infrastructures of power grids, a massive number of data can be available for advanced design of control and management functionalities. The peak load of electric power has reached a high level and the difference between peak and valley load has a continuous growth trend. The limitations of load management mode are becoming increasingly prominent. The data acquisition system developed by using 5G network key technology can provide improved performance of data collection with significantly reduced delay and throughput. It is suitable for various scenarios, reduces the data blind area and has good scalability. The network capabilities of the 5G network, such as ultra-low delay, high bandwidth and high mobility, can meet the rigid needs of the power business. It is the key technology to address the challenge of the current development dilemma of the power Internet of things. It has important application value to study the multi-type energy consumption measurement data acquisition mode and transmission technology, data calculation and processing technology based on 5G digital slicing technology, and realize the efficient management of energy consumption measurement terminal of digital twin platform, accurate portrait of energy consumption behavior and dynamic evaluation of demand side response capability.

With the rapid development of the fifth-generation communication technology 5G, the combination of 5G technology with new generation information technologies such as big data and artificial intelligence will give birth to a series of new applications, new products and new business models. In the field of power engineering, 5G technology will be applied in many links of smart power. With the continuous deployment of metering equipment based on 5G technology, a large number of metering data will be generated in the integrated energy system. In the integrated energy system, various energy sources have the problems of strong uncertainty, different time scales, poor operation specificity and different response speed. Because 5G can customize network slices, Therefore, the data acquisition requirements in corresponding scenarios can be realized according to the time scales of different energy sources, to realize the coordinated and optimal scheduling under high uncertainty in the integrated energy system. At the same time, due to the strong uncertainty of distributed energy and the complexity of time-space coupling of comprehensive energy, it is difficult to accurately model, and the traditional numerical calculation methods are difficult to meet the needs of real-time. Therefore, the high real-time performance of edge computing can be used to process a large number of data measurements. The efficient analysis and processing of energy data on the customer side can promote the active participation of end users, improve the consumption capacity of renewable energy in the integrated energy system, and maximize the comprehensive benefits.

In parallel, artificial intelligence algorithms have been applied to nonlinear fitting and intelligent decision-making. With the reform of energy structure, the new energy industry has been developing rapidly. The high proportion of distributed energy and the large-scale access of electric vehicles make the power grid structure more complex and flexible, with the characteristics of large uncertainty, strong nonlinearity and complex coupling relationship. The power grid presents an intelligent development trend, and its requirements for power system application technology

2473-3636/22/$31.00 ©2022 IEEE
DOI 10.1109/DCABES57229.2022.00041

tend to be efficient, simple and reliable. However, the traditional technology has some problems such as low reliability, lack of long-term verification and unclear mechanism. Therefore, with its advantages and characteristics, artificial intelligence solutions have become a powerful measure to solve complex power system problems and an effective tool to improve the security, reliability and economy of the new generation of power systems.

However, it should be noted that if there is random discrete data missing, it might result in a reduction in signal-to-noise ratio (SNR) that can directly degrade the data-driven decision-making performance. Therefore, it is of paramount importance to improve the data measurement quality through anomalous data detection and recovery for the practical applications of smart grids.

In literature, many studies have been carried out and a set of solutions are available for anomalous data detection (e.g., [1][2]). The work in [3] developed a deep learning based model to detect the known network attacks and improved the data detection performance. As for the recovery of missing data in the power system, it can be regarded as a statistical interpolation problem. The mathematical methods consist of the mean filling method, polynomial interpolation, and *k*-nearest neighbor method (e.g., [4]). These algorithms are simple and easy to implement, but they are very sensitive to the neighboring data in small intervals before and after the missing data, which requires that the dataset itself varies gently. Besides, those methods cannot perceive the features of data from a large scale of time, which turns in poor performance on long-time continuous missing data. In [5], the authors proposed a kind of supervised cascaded denoising convolutional auto-encoders (CDCAE), aiming to accurately recover the missing load data in an electric power system. In [6], the authors proposed a convolutional self-encoding network (DeCS-Net) for HSI denoising. The proposed solution effectively combines the superiority of convolutional neural network (CNN) and auto-encoder (AE) to learn the multi-scale features. In [7], the authors presented a framework for missing data recovery in the context of missing synchrophasor measurements. The authors of the study [8] presented a spectral graph theory-based method for the recovery of missing harmonic data. The proposed method is validated through simulation and field recorded data. This solution is assessed through experiments and the performance has confirmed that it has excellent anti-noise capabilities with low computational complexity.

To the authors' best knowledge, almost all the proposed recovery methods work within one step strategy, that is to say, to address all the missing data without a difference, rather than to recover them hierarchically. This might lead to extra complexity and limited performance of the algorithms. Another phenomenon is that it seems like the detection and recovery problems are always studied separately. But in fact, these two issues can be combined when we design the model.

This work considers the power demand measurement, i.e. the Belgium load data (www.elia.be) as an example to propose a new model to detect and recover the missing data. This work developed a data detection pipeline with the absolute difference sequence and linear correlation to detect the missing mask and grade the mask at three levels according to the severity of missing. Secondly, since there is an evident periodicity in the load data, we can truncate the one-dimensional data into a two-dimensional matrix by proper cut width, and use the two-dimensional structure as data enhancement. Rather than using the feature vector, this work considers these matrices as generalized image, and transfer the original problem into image inpainting, then propose a cascaded convolutional neural network to hierarchically repair the graded missing data at different levels, providing a new inspiration in this field. The following technical contributions are made in this work:

(1) The proposed solution combines the detection and recovery problems to design algorithms systematically. The solution is extensively evaluated and validated through experiments and analysis.

(2) This work adopted the grade on the missing mask and the design of the recovery network as well. In addition, the pure convolutional structure of the network enables the adaptation to data with different periodic characteristics. The training speed is very fast even for the overall training strategy, and the loss function can be converged easily and stably.

(3) The reconstruction from one-dimensional time series to a two-dimensional image is a powerful data enhancement method for the load data in a power system, which can be an interface between the one-dimensional data and convolutional neural network.

The rest of the work is organized as follows: Section II firstly formulates the problem and presents the proposed solution for data detection and missing data recovery. Section III carries out the simulation experiments and presents the numerical results. Finally, the conclusive remarks are given in Section IV.

II. MODELS AND PROPOSED SOLUTION

A. Missing Data Detection

Compared with the problem of missing data recovery, the problem of missing data detection does not seem to get much attention. In the previous research on missing data recovery, scholars usually label the missing mask artificially, and then design algorithms for recovery. However, in a practical situation, how to automatically detect the missing mask is a kind of complementary part of the repair issue. Therefore, we design a missing detection algorithm based on the characteristics of missing data.

In practice, the measurement value at the missing segments is usually not exactly zero, but the noise is distributed near zero. This kind of noise includes not only the normal noise from sampling equipment and transmission channel but also the noise caused by the instantaneous failure which is the reason for data missing as well. To simplify the model, we can reasonably assume that the noise at the missing segments is subjected to the Gaussian distribution with zero mean value and relatively smaller variance than that of the normal data. According to the obtained missing mask, this work can replace the data at the missing segments with the Gaussian noise, with zero mean value, while the variance, also the noise signal power, should be significantly smaller than the power of the original signal. We will temporarily set that to 1% of the original signal power without any missing data, which is

an SNR of 20dB. It should be noted that the noise feature here is unknown in advance, and the detection algorithm needs to compute it automatically.

To learn the noise characteristics of the missing segments, this work firstly needs to detect or locate some missing segments for calculation, and those segments located in this stage should be of high reliability, otherwise, it will affect the subsequent detection performance, as illustrated in Fig. 7

Figure 1. Examples of atypical missing segments with only one or no jump (yellow)

As we can see, the feature of typical missing segments is more obvious than the other one, which is easier to be detected and located. Therefore, we would detect some of them in advance, as the samples to compute the noise at missing segments. For a typical missing segment, we can see that when the curve enters and leaves the missing segment, the change between adjacent sampling points will increase significantly. To visualize this phenomenon, we can verify this by calculating the Absolute Difference Sequence (ADS). As the result, the typical missing segments can be identified by checking the sudden jump in the absolute difference sequence and the linear correlation.

B. Missing Data Recovery

This work design a data enhancement algorithm to reconstruct the one-dimensional data into two-dimensional matrices, which are regarded as "generalized" images. After then we divide the detected missing mask into three grades and use the image processing techniques in the deep learning field to recover the missing data grade by grade. The load data in the power system will change with the patterns of society operation and production. Thus, it will show significant multiple periodicities in days, weeks, quarters, and years. Due to the fact that the power demand data are generally indirectly adjacent, and share very similar patterns as illustrated in Fig. 2.

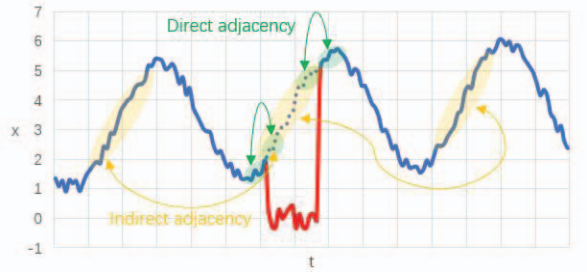

Figure 2. Direct and indirect adjacency relationship in load data

The one-dimensional structure of the measurement data determines that the directly adjacent relationship between the data is visible, while the indirectly adjacent relationship is invisible. To observe this kind of invisible feature, special processing methods, such as Fourier transform, are often needed. This method can analyze the periodic characteristics of data, but also sacrifices the perception of directly adjacent relationships, the occurrence of missing is not necessarily periodic, so it is difficult to restore the missing data at certain locations from the frequency spectrum.

What makes it worse is that Fourier Transform (FT) cannot be directly used for unsteady state signals. Instead, we can only make Short Time Fourier Transform (STFT) for data segments in small window sizes. However, since the window size is fixed, this method won't meet the needs of frequency change for a different signal. After one-dimensional data is truncated and reshaped into two-dimensional square images, the data located in the center of the image will have more available adjacent data than that at the edge of the image, which means the number of available adjacent data is radially attenuated outward, resulting in the recovery of edge data more difficult than that for the central data.

To address the challenge of the uneven distribution of the available adjacent data in the radial direction, some additional redundant data on the four edges around the original image can be rearranged based on the data patterns. Those additional data can supply useful information for the data located on the previous edges. For the convenience of description, we name the original image as the core area and the region of the additional redundant data as the padding area. Besides, we call the core area and the padding area together a padding slice.

III. SIMULATION EXPERIMENTS AND NUMERICAL RESULTS

In experiments, we take the load data of the Belgium grid as the experimental dataset. The preprocessing of data is carried out to normalize the original data values into (-500,0), (-250,250), (0,500) respectively, as the test data for detection.

Since the padding will make no difference to the missing detection problem, we will directly detect the missing mask on the load data. Here we consider two dominant factors in the detection process as the SNR and miss rate. We test our detection algorithm on the data normalized in (-500,0), (-250,250) and (0,500) independently, under miss rate from 10% to 40% and SNR from 15dB to 40dB.

This work evaluates the CPCAE and the improved CPCAE models on different padding depths from 0 to 11, corresponding to the padding ratio from 0% to 23%. During training, we set the batch size as 20 and the initial learning rate as 0.01 which exponentially decreased along the epochs. We will stop the training until the loss function does not decrease any more. The final result is shown in Table 10, where Mean Absolute Error (MAE) is used as a complementary index.

To demonstrate the recovery effect of the improved CPCAE more directly, we randomly chose some results on the test dataset and virtualize them in Fig. 3. To better understand what happens during the training process, we

then virtualize the convolutional kernels in the first layer of the improved CPCAE. There exist obvious horizontal stripes in the convolutional kernels, which are consistent with the characteristics of the "generalized" images. The parameters within these stripes are much lower than the outsides, and the stipes are usually dimmer than in other regions. This is because the missing data are distributed inside these stripes.

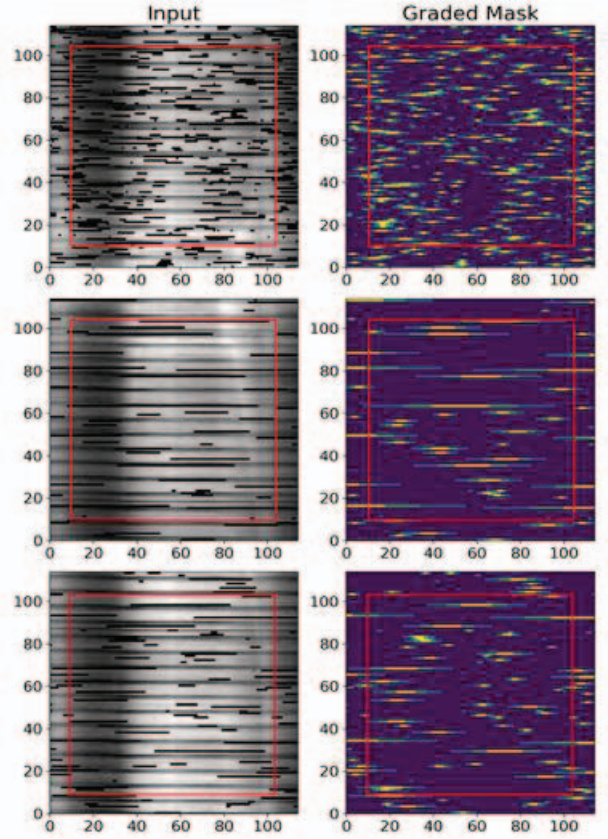

Figure 3. Missing recovery results with core areas inside red rectangles

IV. CONCLUSIVE REMARKS

This paper proposes a missing load data detection and recovery model based on the improved CPCAE. In the detection issue, this work firstly develops an accurate model to describe the missing generation process to generate the missing mask and then combines the absolute difference sequence and the linear correlation as criteria to detect the possible missing segments under different SNR. Based on the detection results, we further divide the detected missing mask into three grades and then reshape the origin one-dimensional data and mask into two-dimensional matrices as a kind of data enhancement. This work considers the constructed matrices as "generalized" images and transforms the problem to recover the missing points in one-dimensional data to probably pinpoint the missing regions in two-dimensional images. Then we intuitively turn to the deep learning technologies on image processing and design an improved CPCAE model to repair the damaged images. As the results on the grid load data show, our detection and recovery model performs well under different missing situations.

For future research, the proposed solution needs to be further evaluated in the more system operational scenarios of power consumption measurements. In addition to the one-dimensional data, further work is required to exploit the extended application of the CPCAE model on anomalous data detection and recovery in other application domains.

ACKNOWLEDGMENTS

This work is supported by the Science and Technology Project of State Grid Zhejiang Electric Power Co., Ltd (5211NB200137).

REFERENCES

[1] K. Singkarn and K. Sinapiromsaran, "C-Anomalous Assemblage Detection to Recognize Outliers Using kth-Nearest Neighbor Distance," 2017 21st International Computer Science and Engineering Conference (ICSEC), 2017, pp. 1-5.

[2] W. Cui and H. Wang, "Anomaly detection and visualization of school electricity consumption data," 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA), 2017, pp. 606-611.

[3] B. Kızıltaş and E. Gül, "Network Anomaly Detection With Convolutional Neural Network Based Auto Encoders," 2020 28th Signal Processing and Communications Applications Conference (SIU), 2020, pp. 1-4.

[4] ENDERS C K. Applied missing data analysis[M]. New York, USA: Guilford Press,2010:22-50.

[5] Y. Chen, Y. Wang and Q. Yang, "Cascaded Denoising Convolutional Auto-Encoders for Automatic Recovery of Missing Time Series Data," 2020 19th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES), 2020, pp. 283-286.

[6] X. Liu, S. Mei, Z. Zhang, Y. Zhang, J. Ji and Q. Du, "Decs-Net: Convolutional Self-Encoding Network for Hyperspectral Image Denoising," IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, 2019, pp. 1951-1954.

[7] P. Gao, M. Wang, S. G. Ghiocel, J. H. Chow, B. Fardanesh and G. Stefopoulos, "Missing Data Recovery by Exploiting Low-Dimensionality in Power System Synchrophasor Measurements," in IEEE Transactions on Power Systems, vol. 31, no. 2, pp. 1006-1013, March 2016.

[8] R. Xu, X. Ma, Y. Wang, Q. Fu, J. Zhao and R. Zhou, "Spectral Graph Theory-based Recovery Method for Missing Harmonic Data," in IEEE Transactions on Power Delivery, doi: 10.1109/TPWRD.2021.3135075.