A Data Encryption and Approximate Recovery Strategy Based on Double Random Sorting and CGAN

Dong Liu[1*], Haoda Wang[1], Quanhong Liu[1] and Yuangang Zhou[1]

[1]College of Electrical and Information Engineering, Hunan University, Changsha, Hunan, 410082, China

Abstract. The measurement data in the power system may be attacked and tampered during the transmission. For this problem, a data encryption and approximate recovery strategy based on double random sorting and CGAN is proposed. The double random sorting encryption algorithm is proposed based on the form of measurement data in plaintext. The randomness of pseudo-random number, random sequence and random sequence insertion will ensure the uncertainty of data block number, sorting order and insertion location of random sequence, which will definitely improve the security of measurement data. Besides, an approximate recovery strategy used to approximately recovery abnormal samples is proposed on the base of CGAN, which can ensure the accuracy of decryption. Finally, the feasibility, and effectiveness of the proposed method are analyzed on wind power and photovoltaic power historical dataset.

1 Introduction

The rapid development of artificial intelligence algorithms and intelligent equipment promotes the development of the power system, which may lead to more potential vulnerabilities in the smart grid and bring great threats to the operation of the power system and the plaintext transmission of data. For example, [1] revealed the threat of network vulnerabilities to the power system, which means that attackers can attack the power system by taking advantage of the potential vulnerability in the cyber network. In addition, some real-time data in the power system is transmitted in plaintext form, like telemetering data. Such unencrypted data will provide attackers with more opportunities [2].

Fortunately, Data encryption technology has been widely used on preventing unknown attacks because of its wide applicability and simple structure. [3] proposed a micro encryption technology combining with the structure characteristics of IEC61850-9-2LE message, which realized the efficient encryption of data. A blockchain-based cloud storage encrypted data with smart contracts was proposed to solve the integrity problems [4]. For long data samples, grouping encryption can be used out to guarantee the security of ciphertext, like Electronic Code Book [5].

In order to successfully attack the power system, attackers will attack the measurement data with low security requirements or uncertain data. Therefore, it is significant to ensure the security of uncertain data. However, the uncertain data in the power system become more difficult to predict with the continuous integration of new energy, especially wind power and photovoltaic (PV) power. The wind power scenarios and PV power scenarios will be used as the input of day-ahead dispatching model [6]. Therefore, if these scenarios are tampered, the security of real-time operation of the power system will be affected. The accuracy of the data can be ensured statistically. For example, a versatile distribution function was proposed, which can fit the data set with uneven tail distribution [7]. [8] proposed a scenario generation method based on conditional generative adversarial networks (CGAN) which is not restricted by dataset. However, the above methods can ensure the quality of the initial data, it can not ensure the security of the data during transmission.

At present, the encryption algorithms are mostly designed for measurement data in the power system, a few encryption methods are suitable for new energy power data. In addition, there are some shortcomings of the traditional encryption algorithms, such as complicated generation of key, so they are not suitable for the encryption of uncertain data. Therefore, as for this problem, we proposed a data encryption and an approximate recovery strategy based on double random sorting and CGAN. The security of cipher will be guarantee by randomness of pseudo-random number, random sequence and the location of random sequence. Besides, the approximate data generated by trained CGAN are used to recovery the detected abnormal data in order to ensure the accuracy of the decrypted data.

*Corresponding author's e-mail: liudyjs@hnu.edu.cn

2 Traditional encryption algorithm and random number generator

2.1Traditional encryption algorithm

Classical encryption algorithms include symmetric encryption algorithm, asymmetric encryption algorithm and message-digest algorithm. The symmetric encryption algorithm is the method that adopts the same key in the processes of encryption and decryption. Common symmetric encryption algorithms include RC5, AES and DES, etc. Due to the efficient encryption speed, symmetric encryption algorithm is generally used in situations where no requirement of key exchange, plaintext data transmission and high requirements of encryption and decryption speed.

The process of asymmetric encryption algorithm is similar to the process of symmetric encryption, and the difference is that different keys are adopted in encryption and decryption, but the two keys are not completely isolated. Classical asymmetric encryption algorithms consist of RSA, RABIN and Elliptic Curve Cryptography, etc. Compared with symmetric encryption algorithm, asymmetric encryption algorithm can achieve a more security of encryption, but it will cost more time. Therefore, it can be used in situations where high requirements of security and key exchange, etc.

Message-digest algorithm is an irreversible encryption algorithm, which converts plaintext data into several fixed short length data segments by specific functions, such as MD5, SHA and Base64 encoding algorithms. Due to the irreversibility of message-digest algorithm, it is commonly used for the integrity and correctness verification of files and data.

2.2Random number generator

Random number consists of real random number and pseudo-random number. The real random numbers can be obtained through the unpredictable physical phenomena, while pseudo-random numbers can be generated by a pre-set method. What is more, the pseudo-random number is similar to real random numbers only when it owns the randomness, unpredictability and non-reproducibility. E.g., when the pseudo-random number is unpredictable, the unknown attacker cannot obtain the formula of the pseudorandom number by analysing large amount of pseudorandom numbers.

3 Conditional Generative Adversarial Network

CGAN is composed of a generator G and a discriminator D. Generator G can capture the distribution characteristics of historical samples and generate false samples based on the random vector and the captured characteristics [9], which are very similar to the real samples. The main function of discriminator D is to distinguish whether the input samples are the real samples or the false samples generated by the generator G. The structure diagram of CGAN is shown in Figure 1.

The training of CGAN can be regarded as a max-min optimization problem. First, G generates the initial false samples according to the random vector Z and the constraint condition C. Second, D will classify real samples X and false samples X'. Third, G and D respectively optimize their network parameters according to the classification results returned by D. Moreover, the parameter of generator G and discriminator D will be optimized alternately until the network attain the Nash equilibrium.

In addition, there are other kinds of GAN developed from the initial GAN, such as DCGAN, WGAN DGAN [20]. Compared with other kinds of GAN, the advantage of CGAN is that it can generate data samples of specified types by setting different condition C and random vector Z. So some special wind power samples or PV power samples can be generated, such as a sample with small gap between its valley and its peak or a sample with a low climbing rate. Therefore, this paper adopts CGAN as the generator of abnormal samples in the proposed data recovery method.

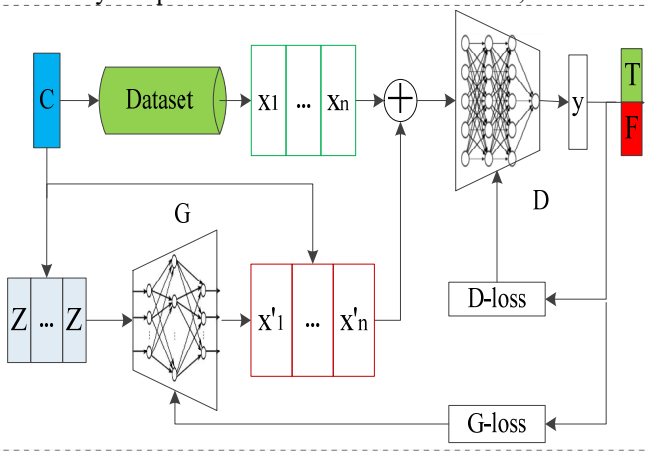

Figure 1. The diagram of the structure of Conditional Generative Adversarial Network

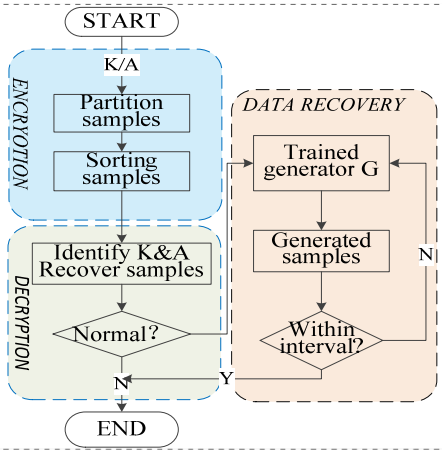

Figure 2. The overall flow of the encryption algorithm

4 Data encryption and approximate recovery strategy

We proposed an encryption method and an approximate recovery strategy based on double random sorting and CGAN, inspired by the algorithm of password based encryption (PBE). Compared with PBE algorithm, our method does not adopt the coordination strategy of "password" and "salt". Instead, we made use of the non-reproducibility of pseudo-random numbers, the random ordering of samples and the random insertion strategy of random sequences can ensure the security of the transmitted data.

4.1Double random sorting encryption method

The double randomness in our method is consistent with the key and its protection policy, which can ensure that the random sequence is not simply identified by an attacker or hacker.

First, a random number K is generated by a pseudo-random number generator, which will be used both as the number of blocks of data and as the insertion position of the random sequence in order to prevent attackers from stealing the features of the data. Secondly, according to K, a random sequence A is generated as the index vector of random sorting, whose randomness reduces the observability of the data. This A is similar to the function of the key, but its generation process is simpler. Finally, the data are randomly sorted according to A. Besides, the A will be scaled appropriately before inserting in the data, so that it owns the same magnitude with the data to avoid being recognized by attackers.

4.2 Approximation recovery strategy based on CGAN

An approximate data recovery strategy is proposed based on the CGAN to recover the identified abnormal data. The trained generator G can generate samples X' within a pre-set confidence interval.

The objective function is equation (1).

$$\min_{G}\max_{D} Loss(G,D) = -\mathrm{E}[D(X|c)] + \mathrm{E}[D(X'|c)] \quad (1)$$

where, $D()$ is the classification result of D. $E()$ is a function used to calculate the mean of the data.

The loss function of G and D is depicted as equation (2) and equation (3) respectively.

$$G\_loss = -\mathrm{E}[D(X'|c)] \quad (2)$$

$$D\_loss = -\mathrm{E}[D(X|c)] + \mathrm{E}[D(X'|c)] \quad (3)$$

The aim of generator G is to generate false samples with the same distribution characteristics of real samples, so it is can not completely reproduce the abnormal samples. Therefore, in this paper, the trained generator G are guided to generate samples within the pre-set c interval and with the pre-set distribution characteristics, which will be used as the approximate recovery samples of abnormal data.

4.3 The data encryption and approximate recovery strategy

Our method can be divided into the above two stages, which is shown in Figure 2. In order to ensure the irreproducibility of generated random sequence, we adopt time seed to ensure that the random number generated by the random number generator has a low repetition.

In addition, the hyperparameters of the trained generator G do not need to be updated in real time, so it will not increase the calculation time of our algorithm. Then, if there emerge some new abnormal samples, these samples will be added to the dataset to build a more balanced dataset, and CGAN will be trained again to update the hyperparameters to improve the accuracy of data recovery.

As for real-time monitoring, if a sample is judged as an abnormal sample, the label corresponding to this sample will be sent to the generator G, which will generate several approximate samples according to the obtained labels and random vectors. Then, determine whether the generated samples are within the pre-set interval. If the samples are within this interval, they will be used as approximate samples of the abnormal sample. Otherwise, generator G will generate new samples.

5 Case study

In this paper, historical wind power samples and historical photovoltaic power samples are from the National Renewable Energy Lab (NREL), and one sample contains of 288 data. The randi function is used to generate pseudo-random number, in which Twister generator is selected as pseudo-random number generator. The allowable range of random number is set as [6, 12]. In addition, the current time is used as the seed of the Twister generator. The randperm function is selected as random sequence generator. This part is run on Matlab2019b on a personal computer. The parameters of CGAN are consistent with those in [14]. The learning ratio is 0.0001, and the confidence interval is set [0.85, 0.95]. This part is run on Python 3.7.4 on another personal computer.

5.1 Data integrity analysis

Encryption algorithms generally guarantee the integrity of data by some special authentication codes. However, the dataset in this paper is the wind power and PV power that allows fault tolerance and deviation, so it is not appropriate to use authentication codes to verify the integrity of samples. Therefore, according to the randomness and uncertainty of samples, this paper define three parameters, including data restoration degree R_{revert}, climbing risk S_{ramp} and peak volatility $V_{variation}$, to evaluate the integrity of data after decryption. The three parameters are formulated as equation (4)-(6).

$$R_{\mathrm{revert}} = (N_{normal}/N_{\mathrm{totall}}) \cdot 100\% \quad (4)$$

$$S_{ramp} = (R_{\max} - r_{\max}) / R_{\max} \quad (5)$$

$$V_{\text{var}iation} = \underset{i\in\Omega}{E}(e_i) \quad (6)$$

Where, N_{normal} is the number of samples whose the decrypted data is consistent with the initial plaintext. N_{total} is the total number of samples. r_{max} is the maximum climbing amount of the decrypted samples. R_{max} is the maximum climbing amount of the historical samples. e_i is the error between the decrypted samples and the historical samples. Ω is the index table of ten moments near the peak point.

The encrypted random sequence of wind power sample is directly inserted without scaling, and the encrypted random number sequence of photovoltaic power samples is multiplied by 0.2 before inserting in order to prevent being attacked easy. First, experiments were carried out on a dataset consisting of 200 samples without noise and 200 samples with noise, and the results are shown in table 1. The Gaussian noise amplitude is [0, 0.3]. It can be seen that as long as the random number and random sequence are successfully located and identified, the decryption can be completed. However, if there is a noise in the samples, the fluctuation of initial samples will be affected, which will affect the accuracy of random sequence recognition. Specifically, the noise can only affect the accuracy of decryption when its amplitude reaches a higher level, especially for samples with low output levels.

Table 1. Integrity analysis of decrypted data

Parameter	R_{revert}(%)	S_{ramp}	$V_{variation}$	Time(s)
Wind power sample	95.0	0.0279	0.4746	0.0079
Wind power samples with noise	94.0	0.1295	0.3796	0.0074
PV power samples	96.5	0.0287	0.0634	0.0109
PV power with samples noise	97.5	0.2049	0.2060	0.0211

5.2 Data encryption efficiency and approximate recovery accuracy analysis

Encryption efficiency is a key parameter in real-time encryption. Therefore, the time of encryption and approximate recovery of samples without random noise are analysed in this part. The time spent on encryption (TE), the time spent on the decryption (TD) and total time (TT) are shown in table 2. Obviously, both processes are not time-consuming, but the encryption cost more time compared with the decryption. The reason is that the identification of random sequence may be more time-consuming.

Table 2. Efficiency analysis of samples encryption

Number of samples	Wind power samples			PV samples		
	TE(s)	TD(s)	TT(s)	TE(s)	TD(s)	TT(s)
200	0.0079	0.0457	0.0268	0.0109	0.0485	0.0297
500	0.0375	0.1782	0.1079	0.0374	0.1394	0.0884
1000	0.1394	0.3498	0.2446	0.0669	0.2462	0.1566
1500	0.1177	0.5886	0.3532	0.0958	0.4896	0.2927
2000	0.1511	0.8050	0.4781	0.1360	0.6387	0.3874

In addition, 576 data collected in two days were concentrated as one sample, and a dataset with 9,464 wind power samples were used to train CGAN for more than 40,000 times. We find that the CGAN network will reach the Nash equilibrium after about 15,000 iterations. The approximate wind power samples generated by the trained generator G, historical wind power samples and their distribution can be seen in Figure 3. It is obvious that the generated approximate samples have same distribution characteristics with initial samples and can recovery the abnormal wind power samples.

5.3 Abnormal sample detection efficiency analysis

In this section, the integrity of samples is no longer verified by the CAPTCHA, but the S_{ramp} and $V_{variation}$ defined in case 1 are used to identify the abnormal samples that could not be decrypted correctly. The experiments are carried out on a dataset consisting of 200 wind power samples with noise and 200 PV power samples with noise. The S_{ramp} and $V_{variation}$ of decrypted samples are shown in Figure 4. We can see that a sample can be judged as an abnormal sample, when its S_{ramp} is greater than 0.7, and approximate samples need to be generated. Moreover, the mutation value of $V_{variation}$ are basically corresponding to the over-limit of S_{ramp}, which can well reflect the existence of abnormal samples. The detection accuracy of abnormal PV power samples is better, which is because the samples are come from the same season. The distribution of PV power samples is relatively uniform and the peak point is relatively fixed, so it can be effectively detected after being attacked.

6 Conclusion

A data encryption and approximate recovery strategy combining double random sorting and CGAN is proposed under the inspiration of the PBE encryption algorithm. The security of the proposed encryption algorithm can be guaranteed by unpredictability and non-reproducibility of pseudo-random number and random sequence, which can reduce the probability of identification of the encryption regulation and increase the security of transformed data. Besides, the complicated calculation of the key is omitted in our method to save the time of encryption and decryption, which can be seen from the simulation results. For the approximate recovery strategy based on CGAN, the trained generator G can efficiently generate specific samples under the guide of the pre-set condition C and the captured distribution characteristics, which can be used to recovery the abnormal samples.

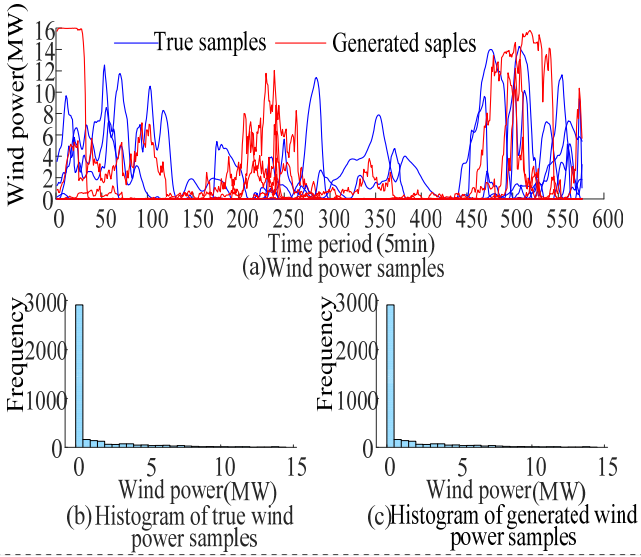

Figure 3. True wind power samples and generated wind power samples (a) and their distribution (b), (c)

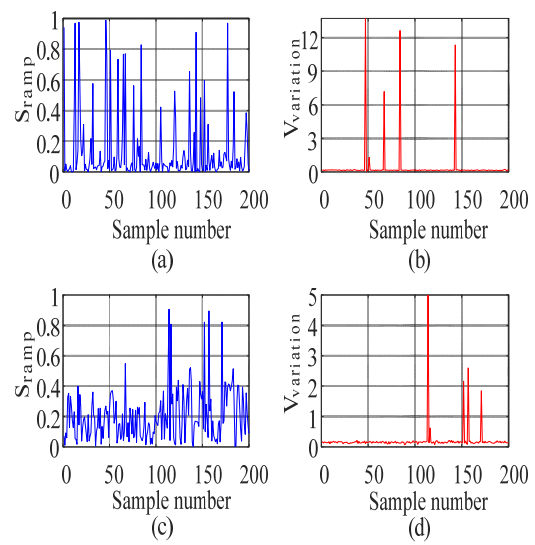

Figure 4. The parameters of wind power samples(a), (b) & PV power samples(c), (d)

Acknowledgments

Thanks for authors in [8] sharing the codes and National Renewable Energy Lab sharing the dataset.

References

1. Musleh A S, Chen G, Dong Z Y. (2020) A Survey on the detection algorithms for false data injection attacks in smart grids. IEEE Transactions on Smart Grid, 11(3): 2218-2234.
2. Song L, Luo Q L, Luo Y, et al. (2004) Encryption on power systems real-time data communication. Automation of Power Systems, 14: 76-81.
3. Wang Z D, Wang G, Li Y C, et al. (2016) IEC 61850-9-2LE message encryption method based on micro-encryption algorithm. Power System Automation, 40(04): 121-127.
4. Liang Y L, Liang J. (2020) Encrypted data sharing scheme in cloud storage based on blockchain. Computer Engineering and Applications, 56(17): 41-47.
5. Wu W L. (2009) Design and Analysis of Block Cipher. Tsinghua University Press, Beijing.
6. Tang C, Xu J, Sun Y, et al. (2018) Look-ahead economic dispatch with adjustable confidence interval based on a truncated versatile distribution model for wind power. IEEE Transactions on Power Systems, 33(2): 1755-1766.
7. Zhang Z S, Sun Y Z, Lin J, et al., (2012) Versatile distribution of wind power output for a given forecast value. In: IEEE Power and Energy Society General Meeting, San Diego, CA, USA. pp. 1-7.
8. Chen Y, Wang Y, Kirschen D S, et al. (2017) Model-free renewable scenario generation using generative adversarial networks. IEEE Transactions on Power Systems, 33(99): 3265-3275.
9. Liu Y P, Xv Z Q, He J H, et al. (2020) Data augmentation method for power transformer fault diagnosis based on conditional Wasserstein generative adversarial network. Power System Technology, 44(04): 1505-1513.