A Survey on Synthetic Data Generation Approaches

Udurume Miracle, Angela Caliwag, Wansu Lim

Department of Aeronautics, Mechanical and Electronic Convergence Engineering
Kumoh National Institute of Technology, Gumi, South Korea
email: wansu.lim@kumoh.ac.kr

Abstract

Experiments based synthetic data has proven to be very effective in simulation models, scientific studies and research purpose. Using several approaches, flexibility and accuracy can be obtained. Synthetic data generation is very useful in cases of limited data availability. This includes cases at which data is available but protected under certain confidentiality by private data collectors and not made available for research purposes. This paper presents a brief survey on synthetic data generation and states some approaches that can be used to carry out synthetic data generation.

Ⅰ. INTRODUCTION

Due to rapid growth in technology and massive increase in connectivity to the internet, a large amount of data is continuously generated [1]. This data is useful in many applications like monitoring and controlling battery pack state of health, prediction of remaining useful life (RUL) of a battery [2]. Due to its wide usage, data has faced some difficulties which can be challenging and painstaking. Despite its wide usage, there are two main challenges of data acquisition: firstly, certain applications require large amount of data which may be difficult to acquire [3]; secondly some data even if available, may still be protected under confidentiality by private data collectors and not freely available for academic or public use [2]. One solution to employ is the use of synthetic data generation (SDG). SDG is the process of generating data artificially for the purpose of preserving privacy, testing systems. It can also be used for creating training data for machine learning algorithm. It can further be used to generate data that meets specific needs and conditions that are not available in existing (real) data. When compared with real data, synthetic data has huge benefits which includes: (i) being able to replicate all important statistical properties of the real data without exposure of real data, (ii) ability to create data to simulate not yet faced problems and (iii) synthetic data also aims at preserving multivariate relationships between variables and not just specific statistical properties.

Synthetic data has found use in many applications which includes: Monitoring, predicting and controlling battery pack state [5], wind turbine fault detection [4], Modeling and evaluating electric vehicle charging sessions [2], probabilistic analysis of hybrid energy systems [7], and so on. In this paper we review three approaches which has been applied for the purpose of generating synthetic data used: Markov chain, ARMA model and GAN. We discuss their applications in SDG.

Ⅱ. SYNTHETIC DATA GENERATION APPROACHES

In this section we briefly explain the approaches used for synthetic data generation.

A. MARKOV CHAIN:

This is a stochastic model used to describe a sequence of possible events in which the probability of each events depends only on the state attained in the previous events. They are based on the traditional probability matrices of various steps [8]. In construction of a Markov model, the state of the system must firstly be determined. In [6], generation of the synthetic data requires two steps; first the EV data set is used to generate transition probability matrices, based on which new current profiles can be generated using the Markov chain propagation (stochastic model). Next is the voltage profile generation which uses two features, the clustering and neural net structure, the output of this network gives the voltage behavior corresponding to the synthetic current profile as shown in Fig. 1.

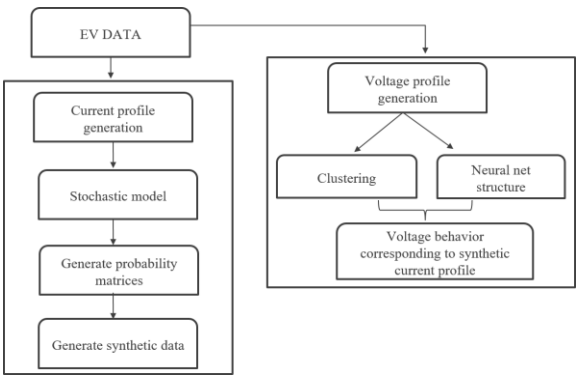

Figure 1: Simple flowchart for ARMA model

Voltage clusters are used as intended targets for neural net structure employed, the clustering algorithm along with the neural net structure is applied to help improve the level of accuracy.

B. ARMA MODEL:

The Autoregressive moving average (ARMA) is used for analyzing a stationary time series. It was put forward by the American statistician G. E. P. Box and

British statistician G. M. Jenkins [9]. In [7] ARMA model is used to model autocorrelation in residue time series (measurements with seasonal trends subtracted) after model is trained over historical data by finding optimal parameters, the combined model is used to generate synthetic data over historical data by finding optimal parameters, the combined model is used to generate synthetic data.

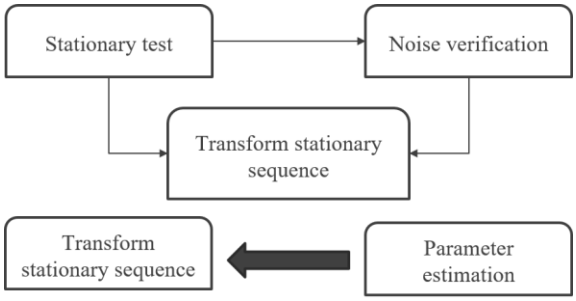

Figure 2: Simple flowchart for ARMA model

C. Generative Adversarial Nets (GAN):

GANs are frameworks that can train deep generating models. There are two networks in a GAN: a generating network G, and a discriminating network D. The generator is used to generate new plausible images while the discriminator classifies images as real (from the domain)

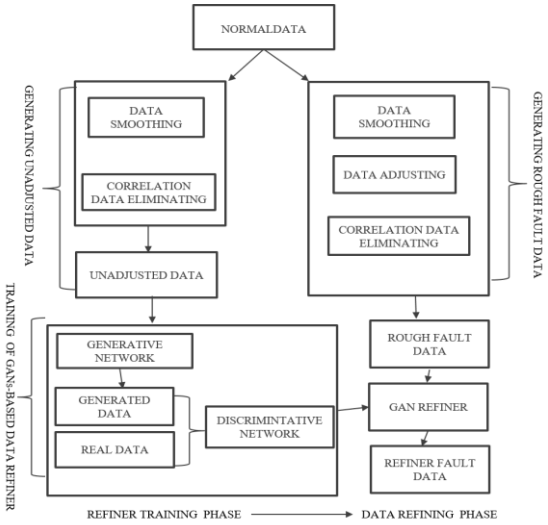

Figure 3: Simple flowchart for GAN model

or fake (generated). In [4], the GAN is used to develop a rough fault data refiner which is used to make the rough data more similar to the real fault data. G is used to synthesize the data resembling real data while D is used to distinguish the synthetic data from the real one. A flowchart of the process is shown Fig. III below.

Ⅲ. CONCLUSION

In this paper a brief survey of some methods used in synthetic data generation was presented. This approaches has been used to solve the challenge of limited data. This next step to this work is to implement the deep learning method used in synthetic data generation in application to real world problems.

ACKNOWLEDGMENT

This work was supported by the Technology development Program (S2829065, S3010704) funded by the Ministry of SMEs and Start-ups (MSS, Korea), and by the Basic Research Program through the National Research Foundation of Korea (NRF) funded by the MSIT (2020R1A4A101777511).

References

[1] P. V. Desai, "A survey on big data applications and challenges," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, pp. 737-740

[2] Lahariya, Manu, Dries Benoit, and Chris Develder. 2020. "Synthetic Data Generator for Electric Vehicle Charging Sessions: Modeling and Evaluation Using Real-World Data." ENERGIES 13 (16). https://doi.org/10.3390/en13164211

[3] Y. Roh, G. Heo and S. E. Whang, "A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective," in IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 4 pp. 1328-1347, 1 April 2021

[4] J. Liu, F. Qu, X. Hong and H. Zhang, "A Small-Sample Wind Turbine Fault Detection Method With Synthetic Fault Data Using Generative Adversarial Nets," in IEEE Transactions on Industrial Informatics, vol. 15, no. 7, July 2019 pp. 3877-3888,

[5] M. Pyne, B. J. Yurkovich and S. Yurkovich, "Generation of Synthetic Battery Data with Capacity Variation," 2019 IEEE Conference on Control Technology and Applications (CCTA), 2019, pp. 476-480

[6] P. Meibom, R. Barth, B. Hasche, H. Brand, C. Weber, and M. O'Malley, "Stochastic optimization model to study the operational impacts of high wind penetrations in ireland," IEEE Trans. Power Syst., vol. 26, no. 3, pp. 1367-1379, 2011

[7] J. Chen, J. S. Kim and C. Rabiti, "Probabilistic analysis of hybrid energy systems using synthetic renewable and load data," 2017 American Control Conference (ACC), 2017, pp. 4723-4728.

[8] F. O. Hocaoglu, O. N. Gerek and M. Kurban, "The Effect of Markov Chain State Size for Synthetic Wind Speed Generation," Proceedings of the 10th International Conference on Probablistic Methods Applied to Power Systems, 2008, pp. 1-4.

[9] Z. Jianyong and W. Cong "Application of ARMA model in ultra-short term prediction of wind power" International computer sciences, 2013