Synthetic data generation: A literature review

Joao Fonseca[1*], Fernando Bacao[1]

[1]NOVA Information Management School, Universidade Nova de Lisboa

*Corresponding Author

Postal Address: NOVA Information Management School, Campus de Campolide, 1070–312 Lisboa, Portugal

Telephone: +351 21 382 8610

This is the abstract for a literature review

1 Introduction

The performance of Machine Learning models is highly dependent on the quality of the training dataset used [1, 2]. The presence of imbalanced and/or small datasets, target labels incorrectly assigned, outliers and high dimensional input spaces reduce the prospects of a successful machine learning (ML) model implementation [2, 3, 4]. In the case of deep learning, for example, these models are often limited by a natural inclination to overfitting, label noise memorization and catastrophic forgetting [5]. Regularization methods are the typical approach to address these problems, but producing robust ML solutions is still a challenge [6].

It is frequently assumed that the training data is sampled from a fixed data source, it is balanced and does not contain label noise. Under these conditions, the resulting ML classifier is expected to achieve good generalization performance [7]. Although, in practical applications, this is rarely the case. When the training data is not representative of the true population, or the model is over-parametrized, it becomes particularly prone to overfitting [8]. Regularization methods attempt to address these limitations. They can be divided into three categories [9]:

1. Output level modifications. Transforms the labels in the training data.

2. Algorithmic level modifications. Modifies the classifier's architecture, loss function or other components in the training procedure.

3. Input level modifications. Modifies the training dataset by expanding it with synthetic data.

The last approach, input level modifications, is known as data augmentation. Data augmentation is used to increase the size and data variability of data in a training dataset, by producing synthetic observations [10, 11]. Since it is applied at the data level, it can be used for various types of problems and classifiers [12]. However, the generation of synthetic data is not only limited to regularization techniques.

Synthetic data generation is also a popular technique to produce synthetic, anonymized versions of datasets [13]. It is considered a good approach to share sensitive data without compromising significantly a given data mining task [14, 15]. Traditional data anonymization techniques, as well as federated learning are two other viable solutions for privacy-preserving data publishing tasks, but contain drawbacks [16]. On the one hand, traditional data anonymization requires domain knowledge, is labor intensive and remains susceptible to disclosure [17]. On the other hand, federated learning is a technically complex task that consists on training ML classifiers on edge devices and aggregating temporarily updated parameters on a centralized server, instead of aggregating the training data [18].

1.1 Contributions

Contributions of this paper:

- Bridge different ML concepts using synthetic data generation in its core (Algorithmic applications + Review of the State-of-the-art).
- List the different synthetic data generation/data augmentation taxonomies and characterize all relevant methods accordingly (Data augmentation taxonomy).
- Discuss the ML techniques in which synthetic data generation/data augmentation is used, beyond regularization (Algorithmic Applications).
- Bring to light the key challenges of synthetic data generation and put forward possible research directions in the future.

1.2 Paper Organization

TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO

2 Synthetic Data Generation

3 Data Augmentation Taxonomy

4 Review of the State-of-the-art

5 Algorithmic applications

References

[1] G. Fenza, M. Gallo, V. Loia, F. Orciuoli, and E. Herrera-Viedma, "Data set quality in machine learning: Consistency measure based on group decision making," *Applied Soft Computing*, vol. 106, p. 107366, 2021.

[2] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 8–12, 2009.

[3] P. Domingos, "A few useful things to know about machine learning," *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[4] S. Salman and X. Liu, "Overfitting mechanism and avoidance in deep neural networks," *arXiv preprint arXiv:1901.06566*, 2019.

[5] Z. Xie, F. He, S. Fu, I. Sato, D. Tao, and M. Sugiyama, "Artificial neural variability for deep learning: On overfitting, noise memorization, and catastrophic forgetting," *Neural computation*, vol. 33, no. 8, pp. 2163–2192, 2021.

[6] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[7] M. Benning and M. Burger, "Modern regularization methods for inverse problems," *Acta Numerica*, vol. 27, pp. 1–111, 2018.

[8] P. L. Bartlett, A. Montanari, and A. Rakhlin, "Deep learning: a statistical viewpoint," *Acta numerica*, vol. 30, pp. 87–201, 2021.

[9] C. F. G. d. Santos and J. P. Papa, "Avoiding overfitting: A survey on regularization methods for convolutional neural networks," *ACM Computing Surveys (CSUR)*, 2022.

[10] D. A. Van Dyk and X.-L. Meng, "The art of data augmentation," *Journal of Computational and Graphical Statistics*, vol. 10, no. 1, pp. 1–50, 2001.

[11] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell, "Understanding data augmentation for classification: when to warp?," in *2016 international conference on digital image computing: techniques and applications (DICTA)*, pp. 1–6, IEEE, 2016.

[12] S. Behpour, K. M. Kitani, and B. D. Ziebart, "Ada: Adversarial data augmentation for object de-

tection," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1243–1252, IEEE, 2019.

[13] F. K. Dankar and M. Ibrahim, "Fake it till you make it: Guidelines for effective synthetic data generation," *Applied Sciences*, vol. 11, no. 5, p. 2158, 2021.

[14] J. Taub, M. Elliot, M. Pampaka, and D. Smith, "Differential correct attribution probability for synthetic data: an exploration," in *International Conference on Privacy in Statistical Databases*, pp. 122–137, Springer, 2018.

[15] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim, "Data synthesis based on generative adversarial networks," *Proceedings of the VLDB Endowment*, vol. 11, no. 10, 2018.

[16] M. Hernandez, G. Epelde, A. Alberdi, R. Cilla, and D. Rankin, "Synthetic data generation for tabular health records: A systematic review," *Neurocomputing*, 2022.

[17] J. P. Reiter, "New approaches to data dissemination: A glimpse into the future (?)," *Chance*, vol. 17, no. 3, pp. 11–15, 2004.

[18] B. Yu, W. Mao, Y. Lv, C. Zhang, and Y. Xie, "A survey on federated learning in data mining," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 12, no. 1, p. e1443, 2022.