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 generation of synthetic data is essential for various domains and tasks. For example, synthetic data is used as a form of regularizing neural network (*i.e.*, data augmentation) **[CITATION]**. One form of anonymizing datasets is via the production of synthetic observations (*i.e.*, synthetic data generation) **[CITATION]**. In settings where only a small portion of training data is labeled, some techniques generate artificial data using both labeled and unlabeled data with a modified loss function to train neural networks (*i.e.*, semi-supervised learning) [1]. In imbalanced learning contexts, synthetic data can be used to balance the target classes' frequencies, reinforcing the learning of minority classes (*i.e.*, oversampling) [2]. Some active learning frameworks use data generation to improve the quality of data selection and classifier training [3]. Other techniques employ data generation to produce deep neural networks without labeled data (*i.e.*, self-supervised learning) [4].

Accordingly, the breadth of these techniques span multiple domains, such as facial recognition [5], Land Use/Land Cover mapping **[CITATION]**, medical image processing **[CITATION]** and Natural Language Processing [6].

According to the domain, the data generation techniques used may vary significantly. Generally speaking, there are data generation mechanisms specific to most domains. However, multiple techniques are independent and can be used for multiple data types, tasks or domains. For example, ...

Depending on the context, evaluating the quality of the generated data is a complex task. For example, for image and time series data, perceptually small changes in the original data can lead to large changes in the euclidean distance [7, 8].

In this literature review, we focus on the Machine Learning perspective of synthetic data, as opposed to the practical perspective. From a practical sense, synthetic data is used as a proxy of real data. It is

assumed to be inaccessible, essential and secondary for different purposes, such as educational, software development, or systems demonstrations [9].

1.1 Contributions

This literature review focuses on the generation mechanisms underlying the different techniques where synthetic data is generated.

We focus particularly on tabular and feature space (*i.e.*, embedded inputs) augmentation given its breadth in scope.

Related literature reviews are mostly focused on specific algorithmic applications, with little to no emphasis on the core generative process. For this reason, these techniques often appear "sandboxed" even though there is a significant overlap between them.

The different taxonomies established are often specific to the technique discussed. However, it is possible to establish a broader taxonomy without giving up on specificity.

With this article, we aim to understand the current research gaps in the different data mining techniques that involve synthetic data generation. We compare the strengths and weaknesses of the models developed within each of these fields. Finally, we identify possible future research directions to address some of the limitations found.

Contributions of this paper are summarized below:

- 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 and consolidate the current data generation mechanisms across the different techniques (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

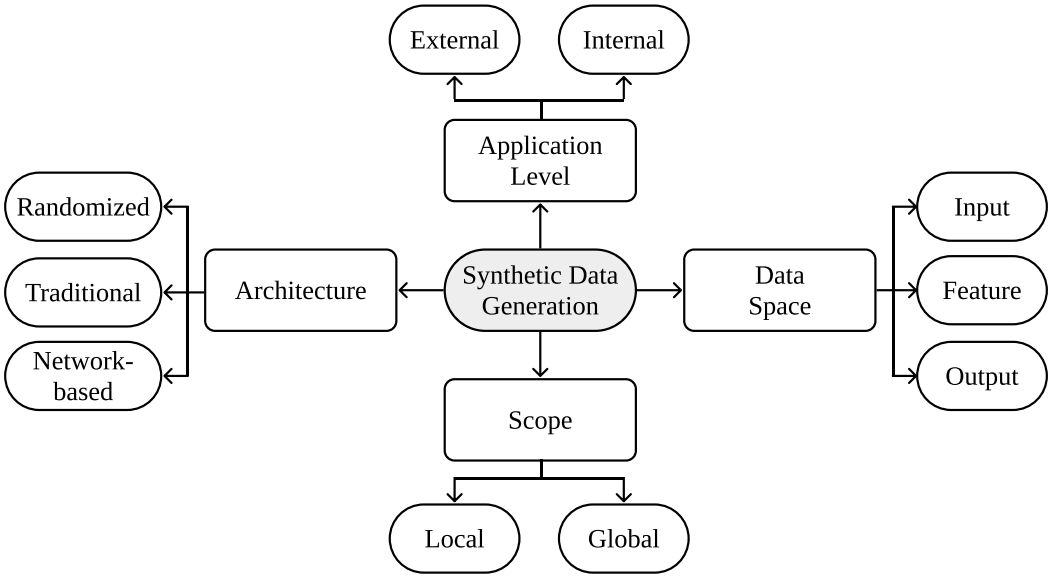

Figure 1: General taxonomy of data generation mechanisms proposed in this paper.

2 Data Generation Taxonomy

Image data augmentation taxonomy [10]

There is a distinction between semantic and traditional image data augmentation [11], also discussed in [12]

Synthetic data generation for medical records taxonomy [13] which is incomplete

Data generation mechanisms can be characterized in 4 properties: Architecture, Application level, Scope and Data space. The overall definition of the proposed taxonomy is shown in Figure 1.

1. Level of application (External or Internal)
2. Scope (Local or Global augmentation)
3. Architectural approach (heuristic, network-based or others)
4. Data space (Input, feature or output). Within feature and output: Domain

3 Synthetic Data Generation

According to [7]. The generation of synthetic data should aim to fulfil the conditions below:

- Privacy preserving.
- Human readable.

- Compact.

4 Data Generation Mechanisms

In this section, we describe some popular domain and data type-specific data generation techniques. For each data type we include a table with related literature reviews specific to different domains.

4.1 Tabular

4.2 Time series

Generative adversarial networks in time series

4.3 Image

Image-specific data generation mechanisms can be further divided into traditional and semantic techniques [11]. Traditional generation techniques comprise simple modifications such as translation, cropping or random erasing [14]. Semantic generation methods involve more complex tasks, such changing colors of specific attributes, backgrounds and visual angles **[CITATION]**.

Data generation by modifying specific attributes in data points with known perturbations [5]. For example, overlaying facial elements into a picture containing a human face (*e.g.*, adding sunglasses and different hairstyles), introducing perturbations in facial landmarks, different illumination and artificial misalignment are different approaches to generate artificial observations for facial recognition.

Generative Adversarial Networks in computer vision [15]

4.4 Text

NLP also benefit from data augmentation [6].

In NLP, there is the challenge of establishing universal rules for text transformations to provide new linguistic patterns [16]

https://github.com/styfeng/DataAug4NLP

5 Algorithmic applications

5.1 Data Privacy

Synthetic data generation is a technique used to produce synthetic, anonymized versions of datasets [17]. It is considered a good approach to share sensitive data without compromising significantly a given data mining task [18, 19]. Traditional data anonymization techniques, as well as federated learning are two other viable solutions for privacy-preserving data publishing tasks, but contain drawbacks [13]. On the one hand, traditional data anonymization requires domain knowledge, is labor intensive and remains susceptible to disclosure [20]. 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 [21]. Although it prevents sharing sensitive data, its applicability is dependent on the task. Dataset anonymization via synthetic data generation attempts to balance disclosure risk and data utility in the final synthetic dataset. The goal is to ensure observations are not identifiable and the relevant data mining tasks are not compromised [22, 23].

The generation of synthetic datasets allow a more flexible approach to the successful implementation of ML tasks. However,

Anonymizing data using synthetic data generation in the financial sector [7].

5.2 Regularization in Supervised Learning

The performance of Machine Learning models is highly dependent on the quality of the training dataset used [24, 25]. 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 [25, 26, 27]. 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 [28]. Regularization methods are the typical approach to address these problems, but producing robust ML solutions is still a challenge [29].

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 [30]. 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 [31]. Regularization methods attempt to address these limitations. They can be divided into three categories [32]:

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 [33, 34]. Since it is applied at the data level, it can be used for various types of problems and classifiers [35].

5.3 Oversampling

5.4 Active Learning

5.5 Semi-supervised Learning

5.6 Self-supervised Learning

6 Discussion

6.1 Main Findings

6.1.1 RQ1: bla bla bla

6.1.2 RQ2: bla bla bla

6.1.3 RQ3: bla bla bla

6.2 Limitations

Research across the different applications appears to be sandboxed even though all techniques integrate synthetic data in its core.

The evaluation of anonymization techniques lack standardized, objective and reliable performance metrics and benchmark datasets to allow an easier comparison across classifiers to evaluate key aspects of data anonymization (resemblance, utility, privacy and performance). These datasets should contain mixed data types (*i.e.*, a combination of categorical, ordinal, continuous and discrete features) and the metrics should evaluate the performance of different data mining tasks along with the anonymization reliability.

Unlike with data privacy solutions, data augmentation techniques generally do not consider the similarity/dissimilarity of synthetic data.

There is a lack of research on oversampling solutions to generate synthetic data with mixed data types and datasets with exclusively non metric features.

oversampling does not seem to be a relevant source of bias in behavioral research and does not appear to have an appreciably different effect on results for directly versus indirectly oversampled variables [36]

6.3 Research directions

Quantifying the quality of the generated data:

1. Realistic
2. Similarity
3. Usefulness (determine purpose and relevant performance metric)
4. Understand the relationship between the 3 factors

7 Conclusions

References

[1] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *International Conference on Learning Representations (ICLR)*, vol. 4, p. 6, 2017.

[2] J. Fonseca, G. Douzas, and F. Bacao, "Improving imbalanced land cover classification with k-means smote: Detecting and oversampling distinctive minority spectral signatures," *Information*, vol. 12, no. 7, p. 266, 2021.

[3] Y.-Y. Kim, K. Song, J. Jang, and I.-C. Moon, "Lada: Look-ahead data acquisition via augmentation for deep active learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22919–22930, 2021.

[4] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21271–21284, 2020.

[5] J.-J. Lv, X.-H. Shao, J.-S. Huang, X.-D. Zhou, and X. Zhou, "Data augmentation for face recognition," *Neurocomputing*, vol. 230, pp. 184–196, 2017.

[6] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy, "A survey of data augmentation approaches for nlp," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, (Online), pp. 968–988, Association for Computational Linguistics, aug 2021.

[7] S. A. Assefa, D. Dervovic, M. Mahfouz, R. E. Tillman, P. Reddy, and M. Veloso, "Generating synthetic data in finance: opportunities, challenges and pitfalls," in *Proceedings of the First ACM International Conference on AI in Finance*, pp. 1–8, 2020.

[8] L. Theis, A. van den Oord, and M. Bethge, "A note on the evaluation of generative models," in *International Conference on Learning Representations (ICLR 2016)*, pp. 1–10, 2016.

[9] M. Mannino and A. Abouzied, "Is this real? generating synthetic data that looks real," in *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, pp. 549–561, 2019.

[10] N. E. Khalifa, M. Loey, and S. Mirjalili, "A comprehensive survey of recent trends in deep learning for digital images augmentation," *Artificial Intelligence Review*, pp. 1–27, 2021.

[11] Y. Wang, G. Huang, S. Song, X. Pan, Y. Xia, and C. Wu, "Regularizing deep networks with semantic data augmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[12] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

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

[14] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 13001–13008, 2020.

[15] Z. Wang, Q. She, and T. E. Ward, "Generative adversarial networks in computer vision: A survey and taxonomy," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–38, 2021.

[16] M. Bayer, M.-A. Kaufhold, B. Buchhold, M. Keller, J. Dallmeyer, and C. Reuter, "Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers," *International Journal of Machine Learning and Cybernetics*, pp. 1–16, 2022.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

[22] K. Singh and L. Batten, "Aggregating privatized medical data for secure querying applications," *Future Generation Computer Systems*, vol. 72, pp. 250–263, 2017.

[23] P. Li, T. Li, H. Ye, J. Li, X. Chen, and Y. Xiang, "Privacy-preserving machine learning with multiple data providers," *Future Generation Computer Systems*, vol. 87, pp. 341–350, 2018.

[24] 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.

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

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

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

[28] 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.

[29] 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.

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

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

[32] 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.

[33] 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.

[34] 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.

[35] S. Behpour, K. M. Kitani, and B. D. Ziebart, "Ada: Adversarial data augmentation for object detection," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1243–1252, IEEE, 2019.

[36] K. K. Hauner, R. E. Zinbarg, and W. Revelle, "A latent variable model approach to estimating systematic bias in the oversampling method," *Behavior Research Methods*, vol. 46, no. 3, pp. 786–797, 2014.