UDC 004.932.2

SYNTHETIC DATA GENERATION FOR MACHINE LEARNING MODEL TESTING

Trofimov V.
Saint Petersburg State University of Information Technologies, Mechanics and Optics
(Saint Petersburg, Russia)

ГЕНЕРАЦИЯ СИНТЕТИЧЕСКИХ ДАННЫХ ДЛЯ ТЕСТИРОВАНИЯ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ

Трофимов В.А.
Санкт-Петербургский государственный университет
информационных технологий, механики и оптики
(Санкт-Петербург, Россия)

Abstract

This article explores the methodologies and applications of synthetic data generation in machine learning (ML). Synthetic data, a pivotal alternative to real-world datasets, addresses challenges such as data biases, privacy concerns, and limited accessibility. The study highlights advanced techniques like generative adversarial networks (GANs), procedural generation, and diffusion models, examining their strengths, weaknesses, and practical applications. A comparative analysis of these methods is presented, along with insights into their integration into ML workflows. The article also discusses the future prospects of synthetic data in emerging fields, including augmented reality, robotics, and digital twins. Ethical considerations, such as data authenticity and potential misuse, are emphasized, advocating for transparent and accountable synthetic data practices. The research underscores synthetic data's transformative potential in enabling robust and scalable machine learning models across industries.

Keywords: synthetic data, machine learning, generative adversarial networks, digital twins, data generation.

Аннотация

В статье рассматриваются методологии и приложения генерации синтетических данных в машинном обучении (МО). Синтетические данные, являющиеся важной альтернативой реальным наборам данных, решают проблемы, связанные с перекосами данных, конфиденциальностью и ограниченной доступностью. Исследование выделяет современные техники, такие как генеративные состязательные сети (GANs), процедурная генерация и модели диффузии, анализируя их преимущества, недостатки и области применения. Представлен сравнительный анализ этих методов, а также даны рекомендации по их интеграции в рабочие процессы МО. В статье также обсуждаются перспективы использования синтетических данных в таких областях, как дополненная реальность, робототехника и цифровые двойники. Особое внимание уделено вопросам этики, таким как подлинность данных и их потенциальное неправомерное использование, подчеркивая важность прозрачности и ответственности. Исследование подчеркивает преобразующий потенциал синтетических данных в создании надежных и масштабируемых моделей машинного обучения для различных отраслей.

Ключевые слова: синтетические данные, машинное обучение, генеративные состязательные сети, цифровые двойники, генерация данных.

Introduction

The increasing reliance on machine learning (ML) models across industries necessitates robust datasets for training and evaluation. However, acquiring real-world datasets that are both extensive and balanced poses significant challenges due to privacy concerns, accessibility issues, and data biases. Synthetic data generation has emerged as a vital alternative, offering solutions to these challenges by creating artificial datasets that simulate real-world scenarios.

This article aims to explore methodologies for generating synthetic data tailored to machine learning applications. It highlights the benefits of synthetic datasets, such as their adaptability, scalability, and ethical considerations, while addressing potential limitations. Through a comprehensive analysis of current techniques and their effectiveness, this work seeks to provide actionable insights for researchers and developers.

The primary objective of this research is to evaluate the efficiency and reliability of synthetic data generation methods in enhancing model performance. By examining case studies and practical implementations, this study identifies key factors contributing to the success of synthetic data in ML model testing.

Main part. Techniques for synthetic data generation

Synthetic data generation encompasses various techniques, including statistical methods, procedural generation, and generative adversarial networks (GANs) [1]. Statistical methods utilize probabilistic models to replicate data distributions, while procedural generation creates data through algorithmic rules. GANs, on the other hand, leverage deep learning architectures to produce highly realistic synthetic data [2].

The table 1 provides a detailed comparison of popular synthetic data generation methods, focusing on key attributes such as data types, scalability, computational complexity, and practical applications. This table outlines the strengths, weaknesses, and practical use cases for different synthetic data generation methods, offering guidance for their appropriate selection.

Table 1

Comparative analysis of synthetic data techniques

Method	Supported data types	Scalability	Computational complexity	Applications	Challenges
Statistical models	Structured	High	Low	Financial data, healthcare analytics	Limited realism for unstructured data
Procedural generation	Semi-structured	Moderate	Moderate	Simulation environments, gaming	Dependence on predefined rules
GANs	Unstructured	Low	High	Image and audio synthesis, NLP tasks	High resource requirements
Variational autoencoders	Structured and Semi-structured	Moderate	High	Dimensionality reduction, anomaly detection	Difficult optimization processes
Diffusion models	Unstructured	Moderate	High	High-quality image and video generation	Computational intensity

Statistical models are particularly suited for structured datasets, such as tabular financial records, while GANs excel in generating complex, unstructured data like images or audio [3]. Procedural generation serves as a versatile technique for semi-structured data, particularly in simulated environments.

Despite its advantages, synthetic data generation faces several challenges, including the risk of overfitting, difficulties in representing rare events, and computational costs. For example, GANs require significant computational resources, making them less accessible for smaller organizations. Additionally, ensuring the diversity and generalizability of synthetic datasets remains a significant hurdle. Synthetic data finds applications in various sectors. In healthcare, synthetic medical records allow researchers to conduct studies without compromising patient privacy [4]. In the automotive industry, companies like Tesla and Waymo use synthetic data to simulate diverse driving scenarios for training autonomous vehicle algorithms. These examples underscore the versatility and practicality of synthetic data.

Practical implementation

The practical implementation of synthetic data generation revolves around its integration into machine learning workflows to achieve more robust and accurate models. The process typically begins with identifying the specific requirements of the dataset, such as its size, structure, and intended use case. Based on these factors, an appropriate generation technique is selected. For instance, GANs are suitable for unstructured data like images, while statistical models work well for structured datasets [5].

One notable application is in model evaluation, where synthetic data is used to test the performance of machine learning models under controlled conditions. By manipulating the properties of the synthetic dataset, such as introducing specific anomalies or rare events, researchers can observe and quantify the model’s responses. This approach is particularly beneficial in areas where real-world data is either unavailable or insufficiently diverse.

In financial fraud detection, for example, synthetic data allows the creation of simulated transactional records that mimic fraudulent and legitimate behaviors [6]. These records enable models to learn distinguishing patterns without accessing sensitive real-world data. Similarly, in autonomous driving, synthetic datasets simulate complex traffic scenarios, including extreme weather conditions and rare road incidents, providing a comprehensive training environment for self-driving algorithms.

The process of synthetic data generation is often iterative, involving multiple cycles of data creation, evaluation, and refinement. Tools like Python’s PyTorch and TensorFlow libraries facilitate this process, providing frameworks for implementing advanced generation methods. For instance, a generative adversarial network can be coded to produce high-quality synthetic images tailored to specific use cases. Additionally, preprocessing steps such as normalization, noise reduction, and augmentation are applied to ensure the dataset’s compatibility with the target ML model.

To further illustrate, consider the pipeline for integrating synthetic data into a machine learning workflow. The pipeline starts with data synthesis, where the raw synthetic dataset is generated. This dataset then undergoes preprocessing to align with the requirements of the target model. Finally, the processed dataset is split into training, validation, and testing subsets, ensuring a comprehensive evaluation of the model’s performance [7].

This workflow is depicted in Figure 1.

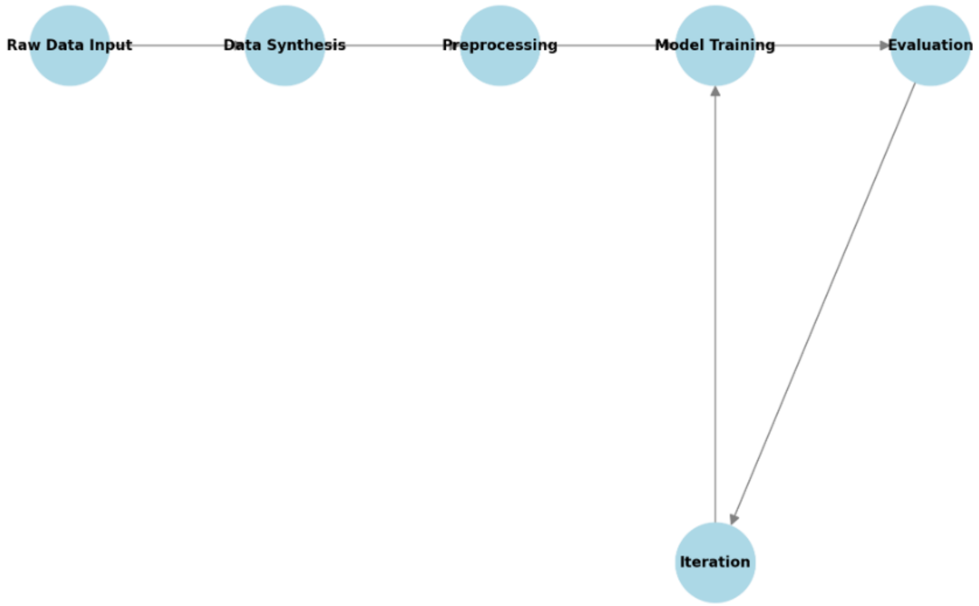

Figure 1. Synthetic data generation pipeline for machine learning applications

This figure illustrates the sequential stages of generating, preprocessing, and integrating synthetic data into machine learning workflows, emphasizing the iterative nature of the process. The iterative nature of synthetic data generation ensures continual improvement in data quality and relevance. This adaptability is crucial in dynamic fields where the requirements of datasets evolve rapidly. By leveraging synthetic data, organizations can mitigate data scarcity and enhance the robustness of their machine learning solutions [8].

Future prospects and ethical considerations

The future of synthetic data generation lies in the development of more advanced and flexible methodologies that address existing limitations while opening new possibilities. Techniques such as diffusion models, capable of producing high-fidelity data for complex applications, exemplify this potential. These models are particularly promising for fields like robotics, where precise environmental simulations are critical for training autonomous systems. Additionally, integration with reinforcement learning frameworks could lead to synthetic data pipelines that adapt dynamically to evolving requirements.

Ethical considerations remain a central focus as synthetic data becomes increasingly prevalent. While it alleviates privacy concerns by reducing dependence on real-world datasets, it introduces challenges related to authenticity and potential misuse [9]. For instance, synthetic data may be leveraged to create misleading content or train malicious algorithms. Addressing these risks requires clear guidelines and robust mechanisms for monitoring and verifying synthetic data applications.

Transparency and accountability are fundamental principles that must guide the implementation of synthetic data. Developers and organizations should prioritize open communication about how synthetic datasets are generated and used, ensuring stakeholders understand their capabilities and limitations. This is especially important in critical domains such as healthcare, where decisions based on synthetic data can have significant real-world consequences.

As advancements continue, interdisciplinary collaboration will play a key role in shaping the ethical and technical landscape of synthetic data generation. By fostering partnerships between data scientists, ethicists, and industry leaders, it will be possible to harness the full potential of synthetic data while mitigating associated risks. The coming years promise a transformative impact, making synthetic data an integral part of the AI development lifecycle and a cornerstone of responsible innovation [10].

Enhanced applications of synthetic data in emerging fields

The increasing demand for innovative applications of synthetic data has expanded its usage into emerging fields such as augmented reality (AR), virtual reality (VR), and advanced robotics. In these domains, synthetic data plays a pivotal role in creating immersive environments and training complex systems. For example, AR and VR systems rely heavily on synthetic datasets to simulate realistic interactions and scenarios, which are challenging to replicate in the real world [11].

A prominent application can be observed in robotics, where synthetic data is employed to train robots for intricate tasks like object manipulation and navigation in unstructured environments. By using synthetic datasets, researchers can expose robots to a diverse range of simulated scenarios, ensuring better generalization and adaptability to real-world conditions. This method has proven especially effective in reducing the cost and time associated with manual data collection. Moreover, synthetic data facilitates advancements in digital twin technology, where virtual replicas of physical systems are developed for monitoring and optimization purposes [12, 13]. Digital twins utilize synthetic datasets to model system behavior under various conditions, enabling predictive maintenance and efficient resource allocation. For instance, the aerospace industry uses digital twins to simulate engine performance, identifying potential issues before they occur.

Conclusion

Synthetic data generation has emerged as a transformative solution to the challenges of acquiring real-world datasets. This technology addresses critical issues such as privacy concerns, data biases, and resource limitations while enabling the creation of versatile and scalable datasets. By leveraging advanced techniques, including generative adversarial networks (GANs) and diffusion models, synthetic data is proving to be a key enabler for advancements in machine learning.

The practical applications of synthetic data span diverse industries, from healthcare to autonomous driving and emerging technologies such as augmented reality and digital twins. Its iterative nature ensures continuous improvements, enabling organizations to refine their models and achieve greater performance under controlled conditions.

As synthetic data continues to evolve, interdisciplinary collaboration and ethical frameworks will play a pivotal role in addressing challenges related to authenticity and misuse. By fostering transparency and accountability, synthetic data generation will remain a cornerstone of responsible AI development, driving innovation across various domains.

References

1. Dankar F.K., Ibrahim M. Fake it till you make it: Guidelines for effective synthetic data generation // Applied Sciences. 2021. Vol. 11. No. 5. P. 2158.
2. Lu Y., Shen M., Wang H., Wang X., van Rechem C., Fu T., Wei W. Machine learning for synthetic data generation: a review // arXiv preprint arXiv:2302.04062. 2023.
3. Das H.P., Tran R., Singh J., Yue X., Tison G., Sangiovanni-Vincentelli A., Spanos C.J. Conditional synthetic data generation for robust machine learning applications with limited pandemic data // Proceedings of the AAAI Conference on Artificial Intelligence. 2022. Vol. 36. No. 11. P. 11792-11800.
4. Rankin D., Black M., Bond R., Wallace J., Mulvenna M., Epelde G. Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for data sharing // JMIR Medical Informatics. 2020. Vol. 8. No. 7. P. e18910.
5. Mendonca S.D.P., Brito Y.P.D.S., Dos Santos C.G.R., Lima R.D.A.D., De Araujo T.D.O., Meiguins B.S. Synthetic datasets generator for testing information visualization and machine learning techniques and tools // IEEE Access. 2020. Vol. 8. P. 82917-82928.
6. Boikov A., Payor V., Savelev R., Kolesnikov A. Synthetic data generation for steel defect detection and classification using deep learning // Symmetry. 2021. Vol. 13. No. 7. P. 1176.
7. Dahmen J., Cook D. SynSys: A synthetic data generation system for healthcare applications // Sensors. 2019. Vol. 19. No. 5. P. 1181.
8. Hittmeir M., Ekelhart A., Mayer R. On the utility of synthetic data: An empirical evaluation on machine learning tasks // Proceedings of the 14th International Conference on Availability, Reliability and Security. 2019. P. 1-6.
9. Kuznetsov I. A. Security and privacy of data in mobile applications developed using machine learning technologies // Cold Science. 2024. No. 2/2024. P. 5-13.

10. Abufadda M., Mansour K. A survey of synthetic data generation for machine learning // 2021 22nd International Arab Conference on Information Technology (ACIT). IEEE, 2021. P. 1-7.

11. Tan C., Behjati R., Arisholm E. A model-based approach to generate dynamic synthetic test data: A conceptual model // 2019 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW). IEEE, 2019. P. 11-14.

12. Kuchin Y.I., Mukhamediev R.I., Yakunin K.O. One method of generating synthetic data to assess the upper limit of machine learning algorithms performance // Cogent Engineering. 2020. Vol. 7. No. 1. P. 1718821.

13. Endres M., Mannarapotta Venugopal A., Tran T.S. Synthetic data generation: A comparative study // Proceedings of the 26th International Database Engineered Applications Symposium. 2022. P. 94-102.