Automated and Manual Synthetic Data Generation for Machine Learning Training

Ms. Molly. M. Mohite[1], Ms. shweta yande[2]
[1,2] *vishnu waman thakur charitable trust*

***Abstract*—This research work provides an in-depth analysis of the accuracy of synthetically generated datasets. The research provides a comprehensive coverage of the application, approaches and basic requirements of the AI-based automated tools in the domain of Machine Learning (ML) with the primary aim of enhancing the ability of the machines to work with human-like intelligence and responsiveness. This research also explores the application of synthetic data generation techniques as a vital component of enhancing the efficacy of model training and alleviating the issues of data scarcity. By the end, readers will have a profound appreciation of the importance and uses of synthetic data, presented in a structured, step-by-step fashion that showcases its crucial contribution to the progress of AI systems .**

***Index Terms*—Synthetic data Generation, Data augmentation, Deep Generative Models ,Tabular GAN**

I. INTRODUCTION:

Although AI has been progressing at a rapid pace in all industries today, we haven't explored its essential components like Natural Language Processing (NLP) , Machine Learning (ML) and Data Science fully. Basic AI is generally understood as technology that behaves like humans, interacts with humans to fix errors or solve problems, with the help of human technical thinking to build systems where machines can work, understand and respond like humans.

Machine learning helps machines to act like humans by training them with relevant datasets and testing their performance with the help of libraries to get accurate results on training and testing data. In some research areas, synthetic data is used as an alternative when data is scarce.

Synthetic data is artificially or manually generated from real data sets, mostly used for solving "edge-case" scenarios. It is particularly useful when real-world data relevant to the task is limited, enabling models to be trained effectively with limited original data.

II. METHOD

You can manually create synthetic data by writing your own rules or examples, or automatically using algorithms and ML pipelines that learn patterns from real data and then generate new records that look similar.

Manual: Manual synthetic data generation generally means rule-based human-designed generation, not fully automated model-based generation. It is especially useful when you want tight control over formats, business rules or edge cases.

Main type

- Random samplingYou use random numbers from specific distributions (normal, uniform, categorical). Simple technique which suits best the basic test datasets without any particular correlation between fields.
- Rule-based generation You describe the rules for field interdependency (e.g., the higher the age, the higher the income). The most straightforward technique with maximal control; used to be the standard way to manually build synthetic data.
- Conditional Generation:You generate data when certain conditionals are met (e.g., within specified age range, valid ID, formatting). This is relevant for datasets that must respect some business rules or constraints of validation.
- Data transformation You manipulate existing data according to your needs through applying certain formula/offset/scaling/masking/reshaping operations to make the new sample. Good if you

need realistic variation while preserving some structure.

- Data augmentationYou transform existing records to produce additional variations (noise injection, rotation, paraphrasing, etc.). Popular technique in image and text processing.
- Simulation-based creation You simulate some complex system or process and generate data based on the results (customers arriving in bank queue, sensor readings from machines). Quite elaborate but still manual technique of generating data

When to use it

Manual synthetic data generation should be considered when high precision, a small to moderate size data set, and adherence to certain business logic rules are required. Manual generation can also come in handy when testing software and generating data while protecting actual personal records.

Automated: Data fabrication through automated methods refers to the creation of realistic datasets using algorithms and pipelines without having to write data records manually. Types of automated synthesis methods include rule-based generation, statistical modeling, simulation-based data generation, and generative modeling techniques such as GANs and VAEs.

Types of automated synthesis methods

- Rule-based generation: Involves creating artificial data such as age, salary, dates, ID numbers, etc., based on logic and rules. This type of data creation technique is usually faster and easier to implement.
- Statistical generation: Analyzes the underlying trends, means, frequencies, correlations in source data and creates new data points that match the same characteristics. Suitable for creating tabular datasets.
- Simulation-based generation: Models processes such as traffic flows, sensor outputs, behavior, and activity, among others, thus creating realistic datasets based on process dynamics. Used for generating realistic datasets from a modeled process.
- GAN-based generation: Applies deep learning methods, including generators and discriminators to create samples that match real data structures. Mostly used for image, text, and tabular datasets
- VAE-based generation: Constructs the probabilistic representations of the source dataset and uses those to sample data points. Suitable for generating datasets where variation needs to be controlled are required.
- Hybrid workflows: Generation, transformation, validation, and scoring processes are all used to evaluate whether the data generated is realistic, relevant, and private. These workflows are typical in enterprise environments.

III. RESULT

Manual limitations :There are two major limitations. First, manual synthetic data generation may become difficult to maintain when applied to big data sets or complicated dependencies between variables. Second, manual techniques produce less realistic results compared to other model-based approaches.

Example: For instance, in generating data about customers, we could consider the following constraints: age ranging between 18 to 80 years, income dependent on age, credit score determined by income, and loan amount determined by credit score.

Automation “when each type fit”: Use rule-based methods for test data, demos, and simple structured fields

- Use statistical or hybrid methods for business tables where preserving correlations matters.
- Use GANs or VAEs when you need more realism and the source data is complex or high-dimensional
- Use simulation when the data must reflect a real process, not just a distribution.

Example: For a hospital dataset, an automated system might generate fake patient ages, diagnoses, and visit dates using statistical rules, then validate that the age ranges, diagnosis frequencies, and field relationships still look realistic while removing any real identifiers. That is the core idea behind privacy-preserving synthetic data

IV. DISCUSSION

Synthetic data generation is an important strategy for model training when real data are limited, sensitive, or expensive to collect. For simple and well-defined tasks, manual synthetic data generation is suitable

because it gives full control over rules, formats, and edge cases, making it useful for small datasets, testing, and validation. However, for large, complex, or high-dimensional datasets, automated synthetic data generation is more effective because it can preserve patterns, correlations, and realism at scale. In practice, rule-based methods are best for structured test data, statistical methods are useful when preserving distributions and correlations matters, simulation is suitable for process-driven data, and deep-learning methods such as GANs and VAEs are preferred when high realism is required. Therefore, the choice of synthetic data type should depend on the goal of the study, the complexity of the data, and the level of realism required for model performance

REFERENCE PAPERS

[1] Veeramachaneni, K. et al. on synthetic data and access to data for machine learning, cited in Coursera's summary article on synthetic datasets.
[2] "Machine Learning for Synthetic Data Generation: A Review" (arXiv review of synthetic data generation methods).
[3] "Synthetic data generation methods in healthcare: A review on open challenges and future research directions" (review article, ScienceDirect).
[4] Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., & Sales, A. P. (2020). Generation and Evaluation of Synthetic Patient Data. (Good healthcare-focused synthetic data paper.)
[5] Torkzadehmahani, R., Kairouz, P., & Paten, B. (2019). DP-CGAN: Differentially Private Synthetic Data and Label Generation. (Useful for privacy-preserving synthetic data discussion.)
[6] Hernandez, M., Epelde, G., Alberdi, A., Cilla, R., & Rankin, D. (2022). Synthetic Data Generation for Machine Learning in Healthcare: Challenges and Opportunities.
(Strong review paper.)
[7] Alaa, A. M., Van Breugel, B., Saveliev, E., & van der Schaar, M. (2022). How Faithful is Your Synthetic Data? Sample-level Metrics for Evaluation. (Useful for discussing synthetic data accuracy and quality evaluation.)