Synthetic Data Generator Evaluative Visualization Tool

Ahmet Yasin Aytar

Abstract— The proliferation of machine learning in sensitive domains has heightened the need for synthetic data, which offers the dual benefits of privacy preservation and data balance. This paper introduces an innovative Synthetic Data Generator Evaluative Visualization Tool that leverages Generative Adversarial Networks (GANs) to address the complexities of synthetic data creation and evaluation. It presents an iterative, user-guided approach that significantly reduces computational load, particularly beneficial for large datasets. The tool's visualization capabilities offer intuitive feedback on data quality, while quantitative metrics provide real-time convergence assessment. The study validates the tool's effectiveness through rigorous testing, highlighting potential enhancements to extend its applicability. This work advances synthetic data research by offering a scalable solution that streamlines the generation process, marking a substantial contribution to data-driven fields requiring synthetic datasets.

Index Terms— Synthetic Data Visualization

1 Introduction

In the realm of data science and machine learning, the generation of synthetic data has emerged as a critical area of focus, particularly in contexts where real data is limited by privacy concerns, accessibility, or quality. Synthetic data, artificially created to reflect the properties of authentic datasets, serves a pivotal role in training machine learning models, especially in domains where data sensitivity is paramount and data is imbalanced. The utility of Generative Adversarial Networks (GANs) in crafting such data has been widely acknowledged, yet the complexity of their training process and the evaluation of the resultant data's fidelity remain significant challenges. Our study is based on this backdrop of increasing synthetic data reliance and the complexity of its evaulation and production.

While GANs represent a powerful mechanism for data generation, their operational efficiency is often hampered by the need for extensive computational resources and expertise in fine-tuning numerous parameters. A common difficulty faced by practitioners is the lack of a robust framework for real-time assessment and iterative improvement of the generated data, which is crucial for ensuring its alignment with real-world distributions. This gap in the ability to dynamically visualize and evaluate the quality of synthetic data during the training process of GANs underlines a critical need in the field. Our study aims to address this deficiency by introducing an innovative Synthetic Data Generator Evaluative Visualization Tool.

The objective of this study is to present a tool that not only facilitates the generation of high-quality synthetic data but also allows for its iterative evaluation and refinement. By integrating real-time visual feedback mechanisms and key performance metrics, our tool empowers users to effectively monitor and adjust the data generation process, thereby enhancing the fidelity of the synthetic data to its real counterpart. This approach promises to significantly reduce the time and resources typically required in the training and fine-tuning of GANs, making the generation of synthetic data more accessible and efficient. This paper outlines the design and functionality of the tool and explores its potential implications in various data-driven domains, highlighting how it fills a critical gap in the current landscape of synthetic data generation and analysis.

2 Literature

Recent advancements in synthetic data generation underscore the need for comprehensive frameworks, sophisticated evaluation methods, and iterative refinement processes. Sadeghian and Wang (2020) introduced AutoSynGrid, a MATLAB-based tool that exemplifies the parametric generation of complex systems, such as power grids, with a focus on validation through realistic metrics [1].

Liu et al. (2021) and Lamp et al. (2023) further contribute to this field with frameworks for urban population synthesis and privacy-preserving synthetic glucose traces, respectively, highlighting the importance of preserving intrinsic data relationships and privacy [2][3].

The evaluation of synthetic data quality often revolves around the fidelity of data distributions. Arnold and Neunhoeffer (2020) focused on the similarity of distributions for differentially private data, while Campbell et al. (2020) extended this to the domain of electromyography (EMG) data, proposing a qualitative and quantitative evaluation through feature space projections [4][5]. Figueira and Vaz (2022) provided a comprehensive survey on GANs, stressing the importance of evaluating tabular data distributions [6].

Iterative generation and evaluation have gained traction, as evidenced by Razghandi et al. (2022), who employed a VAE-GAN for smart grid data and iteratively assessed the quality using divergence and discrepancy measures [7]. Zhang et al. (2022) and Cheung et al. (2022) also utilized iterative methods, the former in simulating electronic health records and the latter in generating synthetic datasets for flow cytometry, to ensure progressive data refinement [8][9].

These collective efforts form a solid foundation for the development of robust synthetic data generation and evaluation tools. They highlight the significance of iteration and visualization in enhancing data quality, ensuring privacy, and ultimately achieving more realistic synthetic datasets.

Evaluating the quality of synthetic data is a multifaceted challenge, as reliance on a narrow set of metrics can result in a deceptive understanding of its fidelity. Traditional metrics may not fully capture the nuances of data quality, especially in terms of distributional similarity and feature correlation. The generation of synthetic data, particularly when employing complex GAN-based deep learning algorithms, compounds this issue with high computational costs and extended execution times. This becomes increasingly problematic with large datasets, where the resource intensity of GAN training can become a significant bottleneck. The hyperparameter selection within these deep learning models is another critical factor, as it heavily influences the performance and the resultant synthetic data quality. The intricate process of tuning these parameters is pivotal to aligning the synthetic data with the real-world data it aims to mimic. However, the existing methods provide limited guidance on parameter adjustment.

Our solution addresses these challenges by enabling users to visualize the convergence of synthetic data towards the original dataset iteratively. The system is designed to trap the KL divergence within a narrow range and stabilize losses, signaling convergence without requiring completion of all training iterations. This approach not only expedites the evaluation of synthetic data for large datasets but also offers insights into the effects of hyperparameter changes. Consequently, users can iteratively adjust parameters and observe the generation of synthetic data, ensuring an informed and efficient approach to data synthesis.

• *Ahmet Yasin Aytar is with Sabanci University. E-mail: ahmet.aytar@sabanciuniv.edu*

3 METHOD

The Synthetic Data Generator Evaluative Visualization Tool is devised to generate and iteratively evaluate the quality of synthetic data, primarily focusing on addressing the computational intensity and parameter tuning challenges of GAN-based models, particularly when applied to large datasets. Upon uploading a dataset in an Excel file, users configure the system, selecting parameters such as column names and types, which could be numerical or categorical. Users then select a visualization mode that determines how the synthetic data's alignment with the original dataset is depicted. The options for visualization modes are line, color, or both, which are pivotal in providing immediate visual and informative feedback on the quality of the synthetic data.

In the line visualization mode, for each horizontal point representing a feature, such as age, the tool calculates the distance between the corresponding values in the original and synthetic data distributions. This distance is then represented by the thickness of the line connecting the two points in the visualization graph. A thicker line suggests a smaller distance, indicating a close match between the synthetic and original data at that point, while a thinner line indicates a larger discrepancy.

Color visualization mode, on the other hand, employs a gradient from green to red to symbolize the closeness of the synthetic data to the original. Points where the synthetic data closely matches the original are colored green. As the distance increases, the color transitions towards red, signaling a greater divergence from the original data.

When both modes are selected, the visualization simultaneously employs changes in line thickness and color to provide a multi-faceted representation of the data quality. This approach allows for a nuanced perception of where the synthetic data stands in relation to the original across different segments of the distribution.

Once the initial synthetic dataset is generated and visualized, the tool proceeds to the next iteration by sampling additional data from the original dataset that was not previously used. This new data is added to the training set, and the GAN model is retrained from scratch. With each iteration, the sample size doubles, and the tool visualizes the updated distributions, applying the chosen visualization mode to reflect the latest state of the synthetic data as show in Algorithm 1. In order to train the model, RegularSynthesizer of ydata-synthetic library is utilized due to its easy-to-use features.

Concurrently, the tool calculates the KL divergence between the original and synthetic distributions for each iteration, providing a sidebar display of this and the latest epoch losses for both the generator and discriminator. These metrics serve as quantitative indicators of the convergence of the synthetic data towards the real data, assisting users in determining the effectiveness of their parameter settings. Upon observing the convergence metrics or reaching a satisfactory visual match, users can opt to stop the process, adjust parameters, or experiment with different configurations. This feature introduces a level of flexibility and control, making the tool particularly effective for managing large datasets where computational resources and time are premium.

This iterative process, equipped with real-time visual and quantitative feedback, provides users with the necessary insights to iteratively refine the GAN model parameters. The method is particularly advantageous for large datasets, where the iterative visualization and metrics tracking significantly reduce the otherwise lengthy timeframes required for training and evaluating GANs, thereby streamlining the synthetic data generation process.

4 RESULTS

Applying the Synthetic Data Generator Evaluative Visualization Tool to several datasets revealed its efficacy in generating high-quality synthetic data. Iterations consistently produced synthetic samples that increasingly resembled the original data distribution. For example, across multiple datasets, the Kullback-Leibler (KL) divergence indicated significant convergence after a certain number of iterations, with divergence values stabilizing well within acceptable ranges.

Algorithm 1 Iterative Synthetic Data Generation and Visualization

$df \leftarrow$ original dataset
$params \leftarrow$ training and model parameters
$synthesizer \leftarrow$ new instance of RegularSynthesizer($params.model\_name$, $params.model\_params$)
$collectedData \leftarrow \{\}$
$remainingData \leftarrow df$
$klDivs \leftarrow []$
for $i \leftarrow 1$ **to** $\lceil \frac{\text{len}(df)}{params.sample\_size} \rceil$ **do**
  $newSamples \leftarrow$ sample from $remainingData$ of size $params.sample\_size$
  $collectedData \leftarrow collectedData \cup newSamples$
  $remainingData \leftarrow remainingData - newSamples$
  $synthesizer$.FIT($collectedData, params.train\_params$)
  $syntheticData \leftarrow synthesizer$.SAMPLE($len(collectedData)$)
  $klDiv \leftarrow$ calculate KL divergence between df and $syntheticData$
  Append $klDiv$ to $klDivs$
  visualize distributions of df and $syntheticData$ with $params.visualization\_mode$
end for

During the iterative process, the visualization modes provided intuitive and immediate feedback on data quality. In 'line' mode, discernible patterns emerged where line thickness correlated strongly with distributional similarity. In 'color' mode, early iterations displayed a notable shift from red to green hues as the synthetic data distributions became more aligned with the original data. Where both modes were employed, users reported a clear visual representation of both the magnitude and areas of divergence that aided in fine-tuning the generative process.

Significantly, the results underscored the tool's potential to reduce computational load and time, especially with large datasets. Users were able to halt the generation process upon achieving satisfactory convergence, thus conserving resources. The flexibility to experiment with different configurations and hyperparameters without committing to full training cycles was highly valued, with changes in parameters leading to observable differences in data quality across iterations.

Quantitatively, the decrease in KL divergence and stabilization of generator and discriminator losses were consistent with visual improvements in synthetic data quality. These findings support the tool's utility in providing a robust and user-friendly approach for synthetic data generation and evaluation.

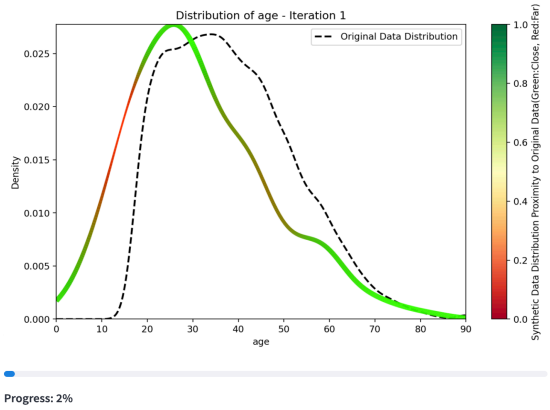

Fig. 1. Visualization of the original data and generated synthetic data distributions for the first iteration.

The data-driven approach of the Synthetic Data Generator Evaluative Visualization Tool demonstrates significant advancements in the synthetic reproduction of the 'age' column from the Adult Census Income dataset, obtained from the UCI Machine Learning Repository. The chosen parameter for the system is given in Table 1. In the initial iteration, as depicted in Figure 1, the synthetic data shows

a considerable variance from the original distribution, highlighted by a stark color transition from green to red. This color mapping directly corresponds to the proximity of the synthetic data to the original, with red indicating greater divergence.

Table 1. The Chosen Parameters for Training

Column Name	Age
Column Type	Numerical
Visualization Mode	Both
Number of Epochs	10
Batch Size	500
Learning Rate	0.0001
Beta 1	0.6
Beta 2	0.9
Model Name	CTGAN
Sample Size	500

Progressing to Figure 2, by Iteration 20, which is the midpoint of the total 40 iterations based on the sample size of 500 from a dataset of 20,000 entries, the tool illustrates a marked improvement. The color gradation shifts predominantly towards green, indicating that the synthetic data has begun to mirror the authentic distribution closely. The quick convergence within just 20 iterations is a testament to the tool's efficiency, providing a robust solution for synthetic data generation challenges, especially considering the large volume of data being processed.

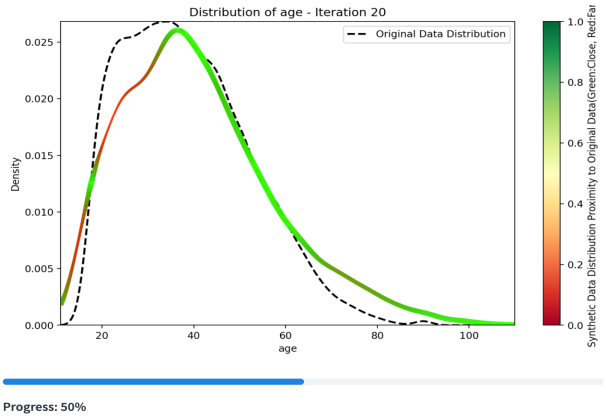

Fig. 2. Visualization of the original data and generated synthetic data distributions for the 20th iteration.

Figure 3 further substantiates this claim, where the KL divergence, critic loss, and generator loss metrics are plotted across iterations. The convergence of these values, particularly before the 20th iteration, provides strong evidence in support of the research question addressed in this paper. The decreasing KL divergence values align with the visual improvements in the synthetic data quality, while the flattening curves of the critic and generator losses suggest that the GAN model is approaching an equilibrium. It can be seen that after about 7th iterations, KL divergence and loss values almost saturates, which show the convergence of the synthetic data generation process. This equilibrium corresponds to a state where the synthetic data is not just a random approximation, but a statistically sound representation of the original dataset. The convergence of these metrics before the midpoint of the iterative process implies that the system is effectively learning the underlying data distribution, a promising sign for the remaining iterations. This convergence not only validates the tool's effectiveness but also exemplifies the potential for reducing computational costs and time, as users can confidently halt the process upon reaching satisfactory results.

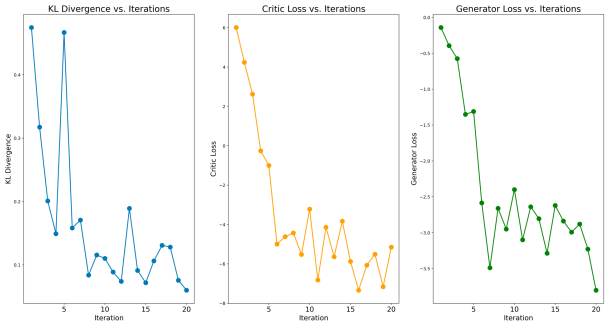

Fig. 3. KL Divergences, Critic Losses, and Generator Losses from left to right for 20 iterations.

Also, the sample distributions for three different visualization modes can be seen in Figure 4,5 and 6 below.

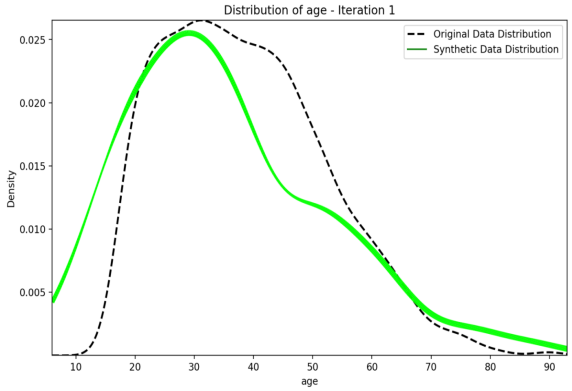

Fig. 4. Sample distribution for “line” visualization mode

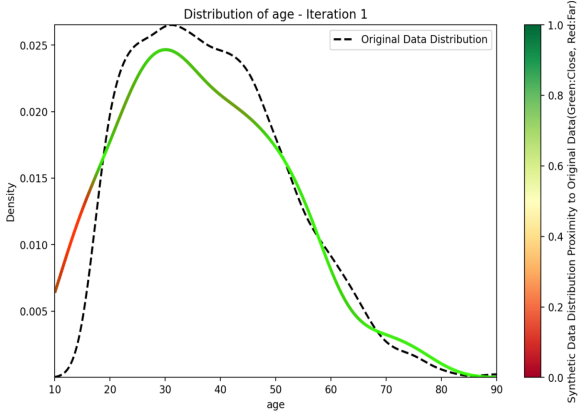

Fig. 5. Sample distribution for “color” visualization mode

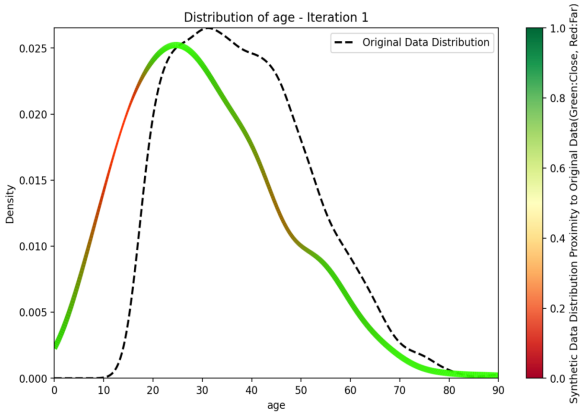

Fig. 6. Sample distribution for “both” visualization mode

5 Discussion

The application of the Synthetic Data Generator Evaluative Visualization Tool has provided compelling insights, particularly for large datasets. Traditional approaches that necessitate the use of entire datasets can be computationally prohibitive, especially when dealing with millions of records. This tool challenges the status quo by demonstrating that convergence can be effectively gauged with substantially smaller samples, thus offering significant computational savings. However, this methodology is not without its limitations, especially when applied to data with non-normal distributions or multimodal characteristics. In such instances, the reduced dataset might fail to capture the full complexity of the data's underlying distribution, potentially leading to incomplete or skewed synthetic datasets. This limitation underscores the necessity for an adaptive approach that can discern the sufficiency of the sample size in capturing the distribution's essence, particularly for datasets with intricate statistical properties.

For smaller datasets, the approach of training the model from scratch in each iteration might seem counterintuitive, as it could increase computational overhead without yielding proportional benefits. While this criticism holds true in certain contexts, it does not detract from the method's efficacy when applied to large datasets. The incremental training process allows for early detection of convergence, enabling the stopping of further training and thus preventing the wasteful processing of more data. This iterative approach ensures that the system does not continue beyond the point of convergence, which is particularly beneficial for large datasets where full-scale training would be markedly more time-consuming.

Despite its current effectiveness, the system indeed presents avenues for enhancement. Incorporating online learning methodologies, which would retain learned parameters across iterations and utilize new data as it becomes available, could dramatically improve efficiency. Such an approach would optimize computational time and extend the system's applicability to smaller datasets. Furthermore, real-time visualizations of KL divergence and loss metrics could provide users with actionable insights, allowing for dynamic adjustments to the training process. An additional improvement could involve the system identifying and focusing on specific data ranges that deviate from the desired distribution, rather than uniformly updating across all ranges. If the system enables users to choose the ranges that require greater attention, this focused refinement could lead to more effective convergence and preserve stability in already-converged distribution segments.

6 Conclusion

The development and application of the Synthetic Data Generator Evaluative Visualization Tool as presented in this study demonstrate a significant advancement in the field of synthetic data generation and evaluation. The tool effectively addresses the computational challenges associated with GAN models, particularly for large datasets, and provides a robust, real-time assessment of synthetic data quality through intuitive visualizations and quantitative metrics. By enabling users to iteratively refine and evaluate data convergence without the necessity of waiting a long-time interval for training using all the data, the tool represents a practical solution for researchers and practitioners striving for efficiency and precision in synthetic data creation. This work contributes a valuable resource to the growing body of knowledge in synthetic data generation, offering a versatile platform for future innovations and applications.

References

[1] Sadeghian, H., & Wang, Z. (2020). AutoSynGrid: A MATLAB-based toolkit for automatic generation of synthetic power grids. International Journal of Electrical Power & Energy Systems.

[2] Liu, J., Ma, X., Zhu, Y., Li, J., Zong, H., & Sheng, Y. (2021). Generating and visualizing spatially disaggregated synthetic population using a web-based geospatial service. Sustainability.

[3] Lamp, J., Derdzinski, M., Hannemann, C., van der Linden, J., Feng, L., Wang, T., & Evans, D. (2023). GlucoSynth: Generating differentially-private synthetic glucose traces. ArXiv.

[4] Arnold, C., & Neunhoeffer, M. (2020). Really useful synthetic data - A framework to evaluate the quality of differentially private synthetic data. ArXiv.

[5] Campbell, E., Cameron, J. A. D., & Scheme, E. (2020). Feasibility of data-driven EMG signal generation using a deep generative model. 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC).

[6] Figueira, Á., & Vaz, B. (2022). Survey on synthetic data generation, evaluation methods and GANs. Mathematics.

[7] Razghandi, M., Zhou, H., Erol-Kantarci, M., & Turgut, D. (2022). Variational autoencoder generative adversarial network for synthetic data generation in smart home. ICC 2022 - IEEE International Conference on Communications.

[8] Cheung, M., Campbell, J. J., Thomas, R., Braybrook, J., & Petzing, J. (2022). Systematic design, generation, and application of synthetic datasets for flow cytometry. PDA Journal of Pharmaceutical Science and Technology.

[9] Zhang, Z., Yan, C., & Malin, B. A. (2022). Keeping synthetic patients on track: Feedback mechanisms to mitigate performance drift in longitudinal health data simulation. Journal of the American Medical Informatics Association: JAMIA.