

Supplementary Material 1. Overview of RTSGAN

Recurrent Time-Series Generative Adversarial Network (RTSGAN) is a novel framework designed to generate high-quality synthetic time-series data. This model addresses challenges such as irregular sampling and missing values, commonly encountered in time-series datasets, particularly in the healthcare domain. By leveraging an encoder-decoder architecture and advanced generation techniques, RTSGAN ensures the production of synthetic datasets that preserve the essential characteristics and temporal dynamics of real-world data (<https://github.com/ratschlab/RGAN>).

Methodology

1. Encoder-Decoder Module: RTSGAN employs an encoder-decoder architecture to map time-series instances into a fixed-dimensional latent space. This module consists of:

- **Encoder:** Compresses input time-series data into a latent representation, capturing essential temporal dynamics and feature distributions.
- **Decoder:** Reconstructs the time-series data from the latent representation, ensuring that the original data characteristics are preserved. The encoder-decoder module ensures an accurate mapping between the input data and its latent representation, maintaining the integrity of temporal patterns.

2. Generation Module: Once the latent space is established, a generation module is trained to produce new latent vectors. These synthetic latent vectors are then passed through the decoder to generate synthetic time-series data that mirror the statistical properties and temporal patterns of the original dataset. This module enables the creation of diverse and realistic synthetic datasets suitable for downstream applications.

3. Handling Missing Values: RTSGAN incorporates mechanisms to address missing data effectively:

- **Observation Embedding Layer:** Utilizes missing patterns as informative features, enabling the model to learn from the data's inherent structure.
- **Decide-and-Generate Decoder:** A specialized decoder design that improves the model's ability to handle missing values while preserving the fidelity of generated data. This approach enhances the quality and realism of synthetic time-series data, even in the presence of missing information.

Benefits of using RTSGAN in generating medical synthetic data

Time-series data are fundamental in healthcare, covering patient vital signs, laboratory results, and other longitudinal measurements. However, these datasets often suffer from irregular sampling and missing values, making analysis and modeling challenging. RTSGAN addresses these issues, offering several key benefits:

- **Data Augmentation for robust model training:** RTSGAN generates diverse synthetic datasets, increasing the size and variability of training data. This improves the performance of machine learning models in predictive tasks, particularly when real-world data are scarce or imbalanced.
- **Privacy Preserving data sharing:** Synthetic datasets generated by RTSGAN do not contain identifiable patient information, enabling data sharing without violating privacy regulations such as the Health Insurance Portability and Accountability Act of 1996 (HIPAA) and the General Data Protection Regulation (GDPR). This facilitates research collaboration and artificial intelligence (AI) model development while reducing the risk of patient re-identification.
- **Enhancing Model Generalization:** AI models trained on limited real-world medical data often suffer from overfitting. By generating diverse synthetic data, RTSGAN helps models generalize better to new and unseen patient cases, leading to more reliable predictions in clinical settings.
- **Cost-Efficient and Scalable Data Generation:** Collecting and labeling real-world medical time-series data is expensive and time-consuming. RTSGAN can generate large-scale synthetic datasets, reducing dependency on real patient data collection.
- **Simulation and Risk Prediction:** Synthetic medical time-series data generated by RTSGAN can be used to simulate different patient trajectories (e.g., disease progression or response to treatment). This can help in predictive modeling and medical simulations.

- **Algorithm Testing and Validation:** Synthetic data offer a risk-free environment for developing, testing, and validating healthcare algorithms before deploying them on real patient data. This ensures algorithm robustness and reliability while mitigating potential risks associated with using untested models in clinical practice.

By addressing challenges in medical time-series data, RTSGAN enhances AI-driven healthcare solutions, supporting more accurate and scalable predictive modeling while maintaining patient privacy and data security.

Experimental Validation

Extensive experiments have demonstrated that RTSGAN outperforms existing generative methods in producing synthetic time-series data. Key metrics for evaluation include:

- **Utility:** Synthetic data generated by RTSGAN retain high utility for downstream classification and prediction tasks.
- **Realism:** Generated data faithfully replicate the intricate temporal relationships and variability present in real datasets.

Conclusion

RTSGAN represents a significant advancement in the generation of synthetic time-series data, particularly for applications in the healthcare domain. By addressing challenges such as irregular sampling and missing values, RTSGAN enables data augmentation, privacy preservation, and algorithm validation, making it an invaluable tool for advancing healthcare research and applications.