Síntese de Séries Temporais utilizando AutoEncoders para Geração de Dados Sintéticos Realistas

Gian Roberto Pereira da Silva[1] and José Luis Seixas Junior[1]

[1]Ciência da Computação – Universidade Estadual do Paraná (UNESPAR)
Apucarana – Paraná

Resumo

Resumo. A geração de dados sintéticos de séries temporais é crucial para mitigar a escassez de dados em domínios industriais. Este trabalho propõe uma metodologia baseada em Autoencoders Convolucionais (1D-CNN) para a síntese de dados de telemetria de propulsores de espaçonaves. Comparado a um baseline denso (MLP), o modelo proposto reduziu o erro de forma (DTW) em aproximadamente 50% e replicou a distribuição estatística com alta fidelidade (Wasserstein de 0.044), validando sua utilidade prática através da métrica TSTR.

1 Introdução

O advento do *Big Data* impulsionou o Aprendizado de Máquina, mas o desempenho desses modelos depende fundamentalmente da quantidade de dados disponíveis. Em domínios críticos, como o aeroespacial, a coleta de dados de falhas ou eventos raros é difícil e custosa [3]. A Geração de Dados Sintéticos (GDS) emerge como uma solução para criar conjuntos de dados anônimos e estatisticamente equivalentes aos reais, permitindo o treinamento robusto de modelos preditivos [7].

O objetivo geral deste trabalho é desenvolver e validar um *pipeline* metodológico para a geração de dados sintéticos de séries temporais. Especificamente, busca-se analisar as limitações de Autoencoders Densos (MLP) e propor uma arquitetura de Autoencoder Convolucional (1D-CNN) capaz de modelar correlações temporais e causais em dados multivalorados de telemetria. A validação é realizada através de um protocolo rigoroso que inclui métricas de forma, distribuição e utilidade.

2 Fundamentação e Metodologia

O conjunto de dados utilizado foi o *Spacecraft Thruster Firing Tests Dataset* [2], caracterizado por sinais de sensores de propulsores de espaçonaves contendo longos períodos de inatividade intercalados por surtos (*bursts*) de alta energia. O pré-processamento evoluiu de uma abordagem unicanal para uma multicanal, integrando as variáveis de comando

(`ton`), força (`thrust`) e fluxo de massa (`mfr`) para preservar a relação física de causa e efeito.

Foram implementadas e comparadas duas arquiteturas de redes neurais. A primeira, um Autoencoder Denso (MLP), serviu como *baseline*, tratando a janela temporal como um vetor achatado [6]. A segunda, a arquitetura proposta, é um Autoencoder Convolucional 1D (1D-CNN), que utiliza filtros deslizantes para capturar padrões locais e invariância à translação [4].

Para o treinamento da CNN, desenvolveu-se uma função de perda híbrida que combina a Entropia Cruzada Binária (para o canal digital `ton`) e o Erro Médio Absoluto (para os canais físicos `thrust` e `mfr`). O protocolo de avaliação adotou a Distância de Wasserstein para medir a similaridade de distribuição [5], o *Dynamic Time Warping* (DTW) para similaridade de forma [1], e a métrica TSTR (*Train-on-Synthetic, Test-on-Real*) para validar a utilidade dos dados em tarefas de regressão.

3 Resultados e Discussão

Os experimentos demonstraram a falha fundamental do modelo *baseline* (MLP). Devido à falta de viés indutivo temporal, o MLP sofreu de colapso de modo, gerando apenas ruído gaussiano em torno da média e falhando em reproduzir os eventos de disparo. Isso foi quantificado por um alto erro de forma (DTW médio de 14.22) e uma alta discrepância na distribuição estatística (Wasserstein de 0.2647).

Em contraste, o modelo proposto (1D-CNN) com perda híbrida obteve sucesso na captura da dinâmica temporal e causal. A Figura 1 ilustra a capacidade do modelo em gerar disparos limpos e alinhados com o comando de ativação.

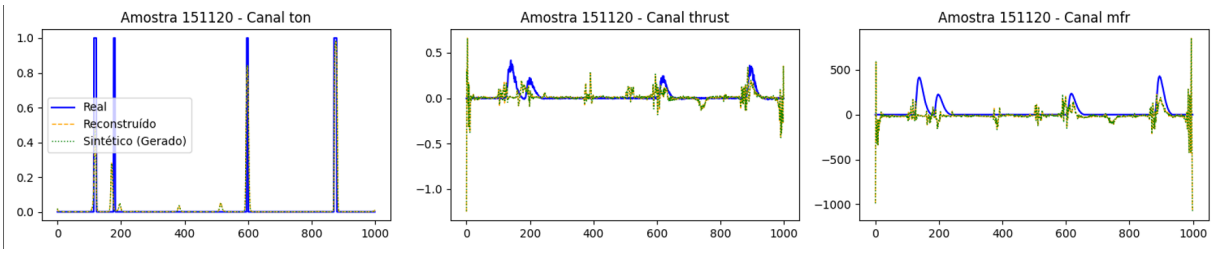

Figura 1: Comparação visual: O modelo 1D-CNN (verde) replica corretamente a forma dos disparos reais (azul), ao contrário do ruído gerado pelo modelo denso.

Quantitativamente, a CNN reduziu a distância de Wasserstein para 0.0447 (uma melhoria de aproximadamente 83%) e o DTW para 7.30 (melhoria de 48%). Na avaliação de utilidade TSTR, o modelo sintético provou ser um gerador de alta fidelidade: um regressor treinado nos dados sintéticos e testado nos reais obteve um desempenho superior ao treinamento com dados reais ruidosos, indicando que o Autoencoder atuou eficazmente também como um filtro de ruído.

4 Conclusão

Este trabalho validou a superioridade das arquiteturas convolucionais (1D-CNN) sobre as densas para a síntese de séries temporais complexas. A introdução da função de perda híbrida foi determinante para modelar sistemas que misturam dados binários e contínuos.

O *pipeline* desenvolvido demonstrou que é possível gerar dados sintéticos que preservam tanto as características estatísticas quanto a dinâmica temporal, viabilizando o uso de dados sintéticos para o treinamento de modelos em cenários de escassez de dados reais.

Referências

[1] Donald J. Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD Workshop*, 1994.

[2] Patrick Fleith. Spacecraft thruster firing tests dataset. `https://www.kaggle.com/datasets/patrickfleith/spacecraft-thruster-firing-tests-dataset`, 2022. Kaggle Dataset.

[3] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[4] Yann LeCun et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[5] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[6] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[7] Jinsung Yoon, Daniel Jarrett, and Mihaela van der Schaar. Time-series generative adversarial networks. In *Advances in Neural Information Processing Systems 32*, 2019.