XTadGAN

Generative Adversarial Networks to Detect Extremely Rare Anomalies

Introduction

Machine learning methods have been widely employed for anomaly detection in time series data, but often struggle to identify rare anomalies in high-dimensional or non-stationary data. Generative Adversarial Networks (GANs) have shown promise in addressing this limitation, but their effectiveness remains a challenge.

This thesis aims to address this issue by devising a method to assess how well different models perform as anomalies become rarer and explore how GANs can be optimized to better detect extremely rare anomalies in time series data.

Main contributions

- **New GAN-Based architectures**. We introduce two new GAN-based architectures to handle rare anomaly scenarios.: **TadGANDT** and **XTadGAN;**
- **Novel Framework for Sensitivity Analysis**. We create a model-agnostic framework for assessing anomaly detection algorithms across a landscape of time series attributes, particularly to varying levels of anomaly rarity;
- **Rarity Sensitivity Analysis**. We conduct a groundbreaking detailed analysis of how different state-of-the-art detection algorithms behave when confronted with variations in anomaly frequency.

XTadGAN

Anomaly Detection in Time Series

- One of the most important data structures in real-world applications
- Immense practical applications
- One of the most researched fields of study using traditional approaches

Extremely Rare Anomalies

- Increases the complexity and difficulty of detection
- Turns an already imbalanced problem in an even more challenging scenario
- Better suited for real-world, often critical, applications

GANs · Generative Adversarial Networks

- One of the "hottest" and more promising fields of study at the moment
- Proven to be very successful in generative contexts (especially images)
- Not much work done leveraging these two fields of study, despite promising results

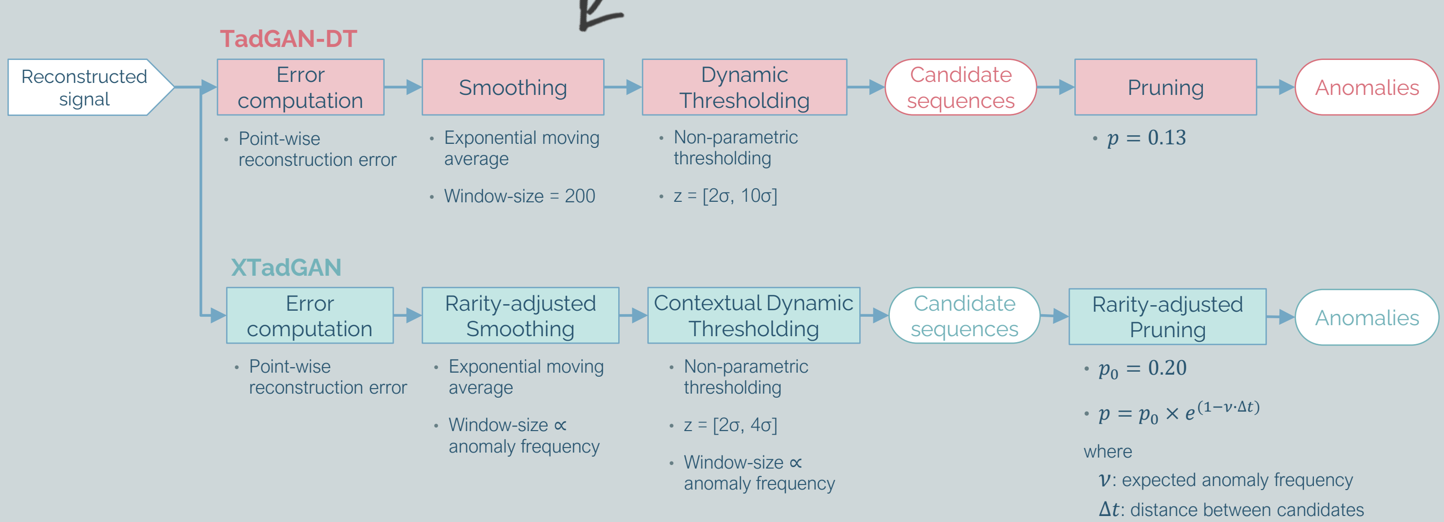
Novel GAN-based Architectures: TadGAN-DT & XTadGAN

TadGAN-DT

Integrates **non-parametric dynamic thresholding and pruning techniques**, enhancing the precision and reliability of anomaly detection in extreme rarity scenarios.

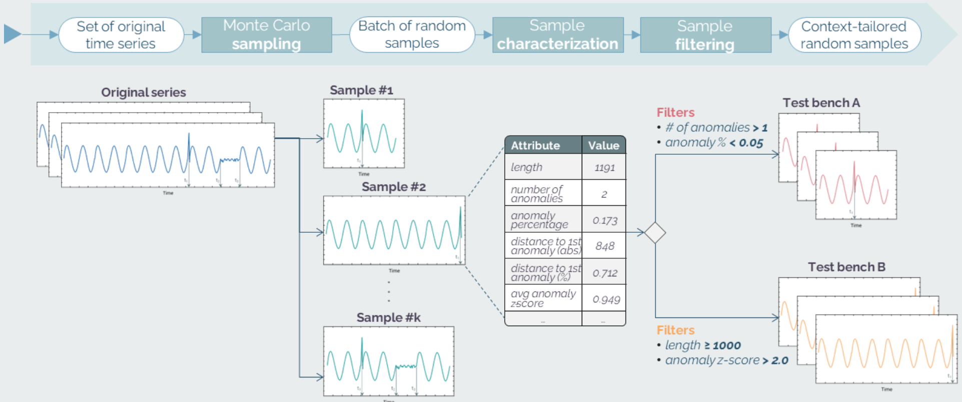
XTadGAN

Leverages *meta-information* regarding expected anomaly frequencies within time series data to establish **rarity-based dynamic thresholding and pruning techniques**, further improving the model's performance in detecting extremely rare anomalies.



New Evaluation Framework: Monte Carlo sampling

A method for **constrained, systematic generation of semi-synthetic** time series. This framework allows the creation of semi-synthetic time series that meet predefined criteria, including constraints on key meta-features related to anomalies

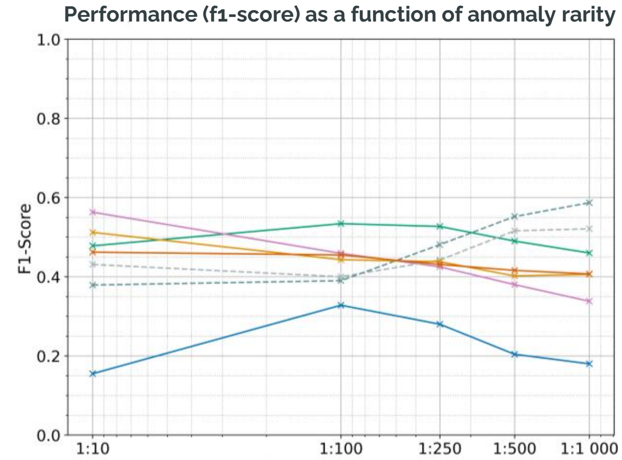


A Monte Carlo approach is employed to create a substantial number of new series by sampling from existing ones. These generated series are then filtered to extract a subset that aligns with the specific requirements/conditions needed for an experiment

Results

Our experimental results demonstrate that **both developed algorithms outperform** other relevant approaches in **rare anomaly detection**.

Performance (f1-score) as a function of anomaly rarity



Algorithm	x_r*	$x_{r \leq 1:500}$
LSTM-DT	0.505	0.475
VAE (LSTM)	0.434	0.412
AE (LSTM)	0.435	0.404
TadGAN	0.429	0.359
ARIMA	0.245	0.192
TadGAN-DT	0.459	0.518
XTadGAN	0.476	**0.570**

* *The **x_r-score** (**rarity spectrum score**) is a newly proposed aggregated measure of performance over the entire spectrum of anomaly levels. It computes the area under the curve for each algorithm and ranges from 0 to 1.*

Thesis · Master in Data Science and Engineering

Nuno Beleza Vasconcelos