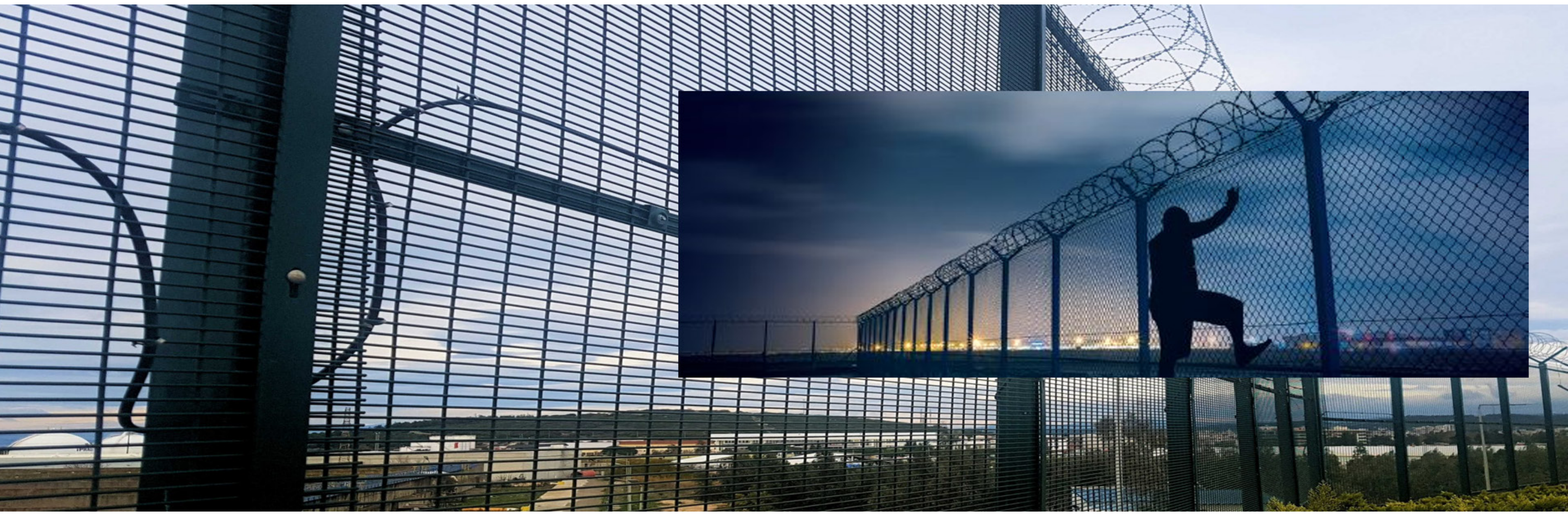
https://doi.org/10.31705/BPRM.v4(2).2024.2



Smart Monitoring
Uncovering Anomalies in Massive Streaming Time Series Data

"Anomaly detection has become a critical component of monitoring systems across various applications in the modern age of digital information and interconnected systems. From detecting infrastructure defects in civil engineering to identifying chemical hazards in environmental engineering, the ability to monitor and detect anomalies in streaming data has a significant impact on safety, efficiency, and operational continuity. With rapid advancements in data collection technology, it has become increasingly common for organizations to rely on sensors to monitor these systems."

Anomaly detection is a multifaceted problem, influenced by how anomalies are defined, the input data type, and the intended output of the detection algorithms. In general, an anomaly refers to a data point, event, or pattern that deviates significantly from the expected behaviour or norm. The landscape of anomaly detection research is diverse, spanning domains such as temporal [1], streaming data [2], and network data [3]. However, one challenge that remains relatively unexplored is detecting anomalous time series within large collections of streaming data—a problem that has significant implications for system monitoring in engineering fields. For example, a sensor cable can be installed on a fence or buried along a facility's perimeter in soil or concrete. Each point along the cable acts as an individual sensor and generates a continuous stream of data over time, forming a time series. Because the cable consists of many sensors, it produces a large collection of time series data. By identifying any unusual time series within this collection, we can effectively detect potential intrusion attempts (Figure 1c).

In this work we propose a robust framework for detecting anomalies within a large collection of streaming time series data. By integrating statistical methods like Extreme Value Theory (EVT), which focuses on predicting the probability of rare and extreme events, and feature-based representations, our approach enables early and accurate anomaly detection across diverse applications.

Types of anomalies in temporal data

In anomaly detection, particularly with time series data, the problem can be divided into three main categories (Figure 1).

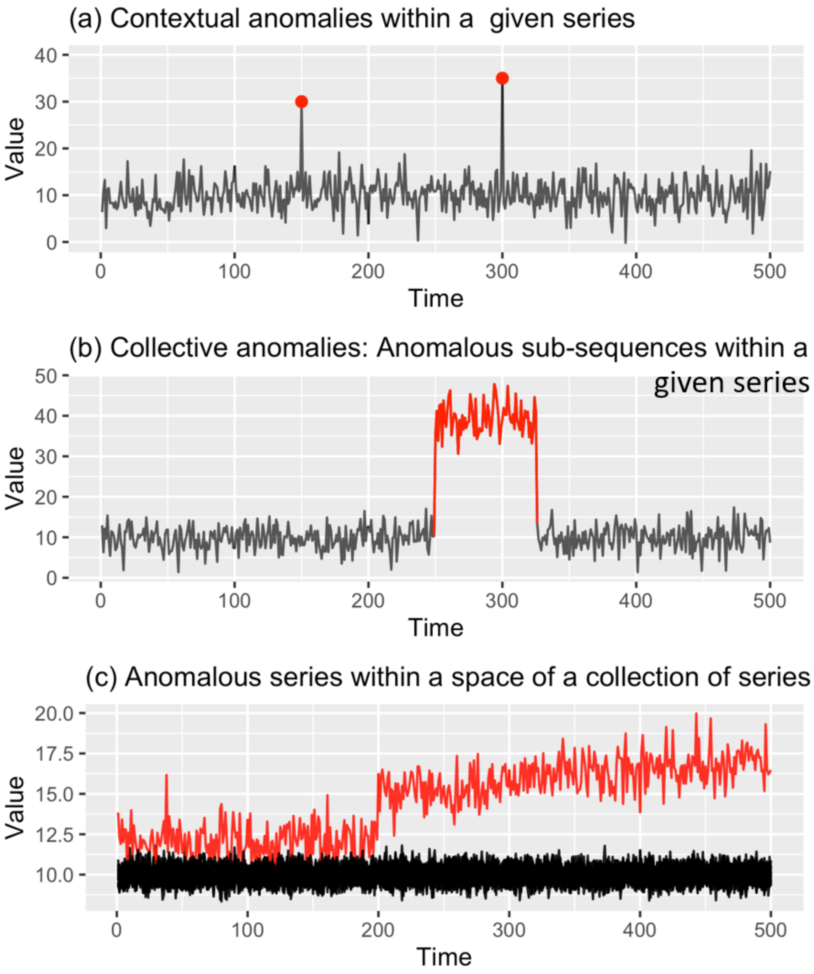


Figure 1: Different types of anomalies in temporal data

a) Contextual anomalies: These are individual observations within a time series that deviate from the expected norm. They are relatively easy to detect, as they appear as outliers compared to surrounding data points.

b) Collective anomalies: In this case, a segment or subsequence of the time series significantly deviates from the rest. Detecting such anomalies requires more advanced methods, as the focus shifts from individual data points to patterns in the sequence.

c) Anomalous Series in a collection of series: The article's primary focus is identifying entire time series that behave anomalously compared to a collection of other time series. This problem becomes more complex when the data is streaming, non-stationary, and noisy, which is often the case in real-world monitoring systems.

Challenges of Streaming Data

The transition from batch processing to streaming data presents significant challenges for anomaly detection. Traditional batch processing assumes that the entire dataset is available before analysis, allowing for detailed inspection of all potential anomalies. In contrast, streaming data evolves in real time, bringing complexities such as high volume, velocity, noise, and concept drift (non-stationarity). These factors necessitate adaptive algorithms capable of distinguishing true anomalies from typical behaviour as the data arrives.

Anomaly Detection in a Collection of Time Series

Many engineering system monitoring problems can be reframed as the detection of anomaly series in a collection of time series. In structural health monitoring, for example, sensors attached to a bridge generate time series data, such as vibration measurements. Anomalies in these time series could indicate rust or cracks in the structure. In mechanical engineering, condition monitoring of multiple machines generates time series for parameters like pressure, temperature, and vibration. Identifying anomalies in these series helps detect malfunctioning machines. Environmental monitoring, such as air quality, soil moisture, and water quality systems, also produces time series data. Anomalous series here can highlight environmental issues.

In electrical engineering, large networks of sensors generate millions of time series, and identifying anomalies can indicate faults in the system. Similarly, railway tracking systems and gas or oil

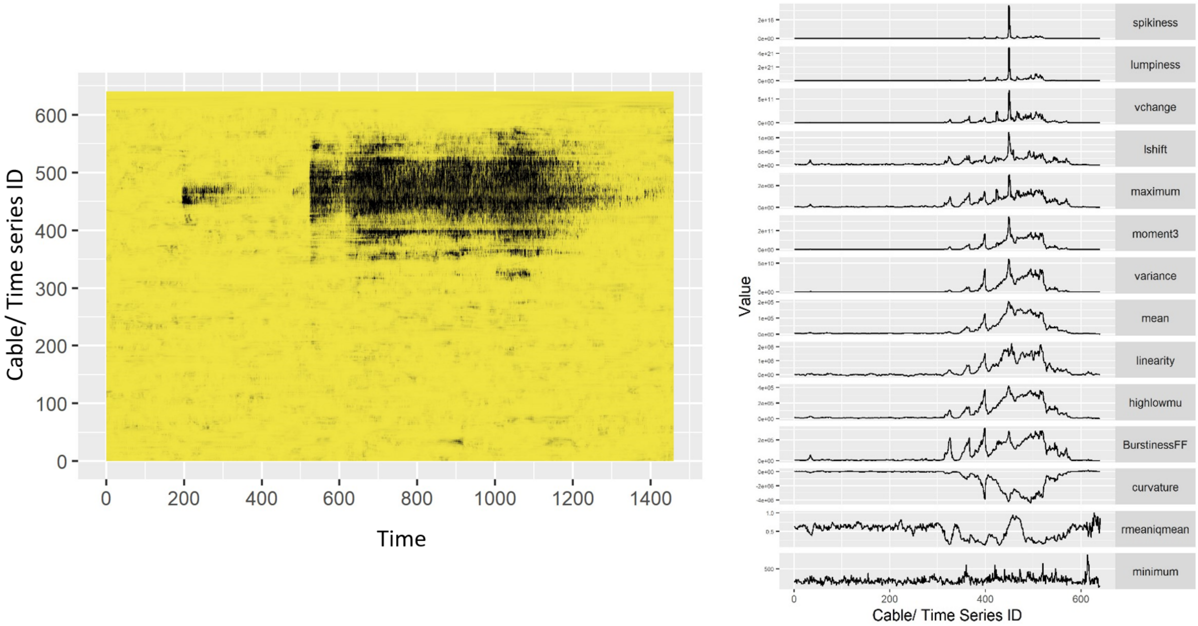


Figure 2: Feature-based representation of the time series

pipelines with sensor cables along their length generate time series, where anomalies reveal track defects or pipeline leaks. In surveillance, fiber optic cables along fences generate time series, and anomalies may signal security breaches.

Across all these fields, identifying anomalies within a large collection of time series can help detect critical issues early, leading to timely maintenance and prevention of failures.

Feature-based Representation of Time Series

Instead of comparing time series point by point, our proposed approach extracts statistical features from each time series in the collection, capturing dynamic properties such as mean, variance, and autocorrelation (Figure 2). This feature extraction step transforms the large collection of time series into a high-dimensional dataset, where each point in the high-dimensional space corresponds to a single time series in the collection. By transforming the time series into feature vectors, algorithms can compare different series more efficiently. This method is particularly useful when dealing with non-stationary or multi-length time series, as it reduces the dimensionality of the problem while retaining important information.

Smart Monitoring Approach

In this work, we propose a novel framework to detect anomalous series within large collections of time series data, especially in the challenging context of non-stationary streaming data [4]. In this proposed framework, we defined an anomaly as an observation that is highly unlikely based on the recent distribution of the system's typical behaviour. The framework is designed to address two critical problems in anomaly detection for time series: detecting anomalies in large collection of time series and identifying changes of the system's typical behaviour, commonly referred to as "concept drift" in a machine learning context.

Firstly, we propose a framework that utilizes extreme value theory to establish a boundary for the system's typical behaviour. This is done in the high-dimensional feature space extracted from the original time series collection. To deal with the streaming data, a rolling window of fixed length is used. This allows the detection of anomalies as they occur without needing to wait for the entire dataset to be available, which is particularly useful for streaming data where real-time decision-making is crucial. The extreme value theory models the tail behaviour of distributions, which makes it effective for flagging rare and extreme deviations from typical behaviour in the system.

Secondly, we introduce a novel approach for detecting non-stationarity (or concept drift) in streaming data. Non-stationarity occurs when the statistical properties of the target variable change over time, making previously learned models unsuitable for the new typical behaviour. In our approach, the algorithm adapts by continuously learning from the incoming data and updating its understanding of "typical" behaviour. The framework monitors changes in the typical behaviour of a given system by conducting density-based comparisons to detect significant shifts in the distribution in the high-dimensional feature space.

Through extensive experiments on both synthetic and real-world datasets, we demonstrate the wide applicability of our framework. The proposed method is resilient to noisy and non-stationary data, making it suitable for a broad range of applications, including industrial monitoring, environmental data analysis, and fault detection in sensor networks. The framework's scalability ensures it can handle large datasets while maintaining high detection accuracy. In many of our applications, we consistently achieve high levels of overall optimized precision while effectively minimizing false detection rates. For an in-depth exploration of our methodology and findings, we invite you to refer to the original paper [4].

The implementation of this approach is available as an open-source package called oddstream in the R programming environment. Oddstream is designed to be user-friendly and efficient, offering an accessible tool for researchers and practitioners to apply the proposed anomaly detection techniques across various domains.

> “Many applications can be viewed as finding unusual patterns in a collection of time series. Our framework not only spots the anomalies in real-time but also adapts to changing data, making it a strong solution for monitoring non-stationary streaming time series”

The rise of large-scale sensor networks and real-time data collection has made anomaly detection an essential tool for system monitoring in various engineering fields. While detecting anomalies within a large collection of time series presents unique challenges, our proposed framework offers a solution that is adaptable, efficient, and scalable for real-world applications.

From structural health monitoring to environmental protection, this approach has the potential to revolutionize how we identify and address anomalies in massive streaming time series data. As the need for smart monitoring continues to grow, advances in anomaly detection will play a pivotal role in ensuring system integrity, safety, and efficiency across industries.

References

[1] Z. Zamanzadeh Darban, G. I. Webb, S. Pan, C. Aggarwal, and M. Salehi, “Deep learning for time series anomaly detection: A survey,” ACM Computing Surveys, 2022.
[2] R. A. Habeeb, F. A. Ariyaluran, A. Nasaruddin, A. Gani, I. A. T. Hashem, E. Ahmed, and M. Imran, “Real-time big data processing for anomaly detection: A survey,” International Journal of Information Management, vol. 45, pp. 289-307, 2019.
[3] S. Ranshous, S. Shen, D. Koutra, S. Harenberg, C. Faloutsos, and N. F. Samatova, “Anomaly detection in dynamic networks: A survey,” Wiley Interdisciplinary Reviews: Computational Statistics, vol. 7, no. 3, pp. 223-247, 2015
[4] P. D. Talagala, R. J. Hyndman, K. Smith-Miles, S. Kandanaarachchi, and M. A. Munoz, “Anomaly detection in streaming nonstationary temporal data,” Journal of Computational and Graphical Statistics, vol. 29, no. 1, pp. 13-27, 2020.

Article by

Priyanga Talagala
Department of Computational Mathematics, Faculty of Information Technology, University of Moratuwa, Sri Lanka