Tools for Advanced Time Series Analytics: Enabling the Future

Nesime Tatbul
Intel Labs and MIT
tatbul@csail.mit.edu

It is widely anticipated that time series data will dominate our future. From autonomous driving to industrial automation, the inevitable rise of the Internet of Things (IoT) has already exposed us to billions of data sources generating time-varying data at unprecedented complexity and scale. The opportunity to utilize this data reliably and efficiently to the benefit of the society critically depends on equipping our computing systems and their users with the right tools and capabilities.

A key capability that lies at the core of managing time series data is the identification of patterns that signify “interesting" phenomena, such as *anomalies*. Anomalies are patterns in data that do not conform to expected (or, normal) behavior [1]. Detecting anomalies before, after, or as they occur not only finds use in many mission-critical domains, but also empowers systems and users with the ability to cope with large data volumes by guiding attention and resources to information that matters the most.

Although anomaly detection has been actively studied for many decades, there are several challenges to be overcome before it can become practical in real-world deployments of time series applications, including the following:

Anomaly detection is a highly domain-specific problem. What constitutes an anomaly changes greatly from one application domain to another. This bears the need to develop, compare, and choose from a variety of models and algorithms. Furthermore, domains may have different preferences for accuracy or performance tradeoffs (e.g., tolerance to different types of classification errors).

Time series anomalies can be complex. They typically fall under the category of collective and contextual anomalies (as opposed to simpler, point anomalies) [1]. These anomalies may expose themselves at different time granularities or aggregations. Furthermore, anomalous patterns may vary over time due to temporal and dynamic nature of the data. Some time series datasets may even consist of multiple, potentially correlated attributes, requiring analyzing anomalies in a multi-variate fashion.

Training data can be unavailable or noisy. In most domains, anomalies are relatively rare or unique events, making it challenging to collect anomalous datasets. Often, training datasets capture a small number of anomalous patterns, which may or may not be properly labeled. In IoT, presence of noisy sensor data makes this situation even worse. As a result, labels may be wrong, incomplete, ambiguous, or missing all together. This may misguide the training process, leading to inaccurate anomaly predictions.

In order to deal with challenges like the above, we believe that it is critical to build tools that will enable data scientists to productively develop and interact with time series anomaly detection systems. This talk will give examples from our own ongoing research in this direction.

Accuracy Evaluation. We have developed a customizable scoring model for range-based anomaly detection, which extends the classical precision and recall model to time series data [5]. Our model provides tunable parameters to capture domain-specific preferences, such as early vs. late detection.

Zero-Positive Training. We have proposed a novel machine learning paradigm for time series anomaly detection, which trains its predictive model based purely on non-anomalous datasets (hence, the term “zero-positive”) [3]. This enables anomalies to be detected even in absence of large, anomalous training data.

Visual Exploration. We have built an interative tool for visually analyzing and experimenting with results of anomaly detectors [2]. Our tool offers a rich set of interaction features (e.g., comparative analysis, what-if testing), designed to assist data scientists in gaining insights about anomalous patterns and what real-world phenomena they may correspond to.

Scaling these tools to work with high-volume and high-velocity datasets also reveals interesting challenges for data management, which we have been broadly exploring within the context of our Metronome Project [4].

REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar. Anomaly Detection: A Survey. *ACM Computing Surveys*, 41(3):15:1–15:58, July 2009.

[2] P. Eichmann, F. Solleza, N. Tatbul, and S. Zdonik. Metro-Viz: Visual Exploration of Time Series Anomalies, 2018. *Under submission.*

[3] T. J. Lee, J. Gottschlich, N. Tatbul, E. Metcalf, and S. Zdonik. Greenhouse: A Zero-Positive Machine Learning System for Time-Series Anomaly Detection. In *Inaugural Conference on Systems and Machine Learning (SysML'18)*, Stanford, CA, February 2018.

[4] J. Meehan, C. Aslantas, S. Zdonik, N. Tatbul, and J. Du. Data Ingestion for the Connected World. In *8th Biennial Conference on Innovative Data Systems Research (CIDR'17)*, Chaminade, CA, January 2017.

[5] N. Tatbul, T. J. Lee, S. Zdonik, M. Alam, and J. Gottschlich. Precision and Recall for Time Series. In *32nd Annual Conference on Neural Information Processing Systems (NeurIPS'18)*, Montreal, Canada, December 2018. https://github.com/IntelLabs/TSAD-Evaluator.

9th Biennial Conference on Innovative Data Systems Research (CIDR '19)
January 13-16, 2019, Asilomar, California, USA.