Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress (Extended Abstract)

Renjie Wu
Computer Science & Engineering Department
University of California, Riverside
rwu034@ucr.edu

Eamonn J. Keogh
Computer Science & Engineering Department
University of California, Riverside
eamonn@cs.ucr.edu

I. Introduction

Most of the time series anomaly detection papers tested on a handful of popular benchmark datasets, created by Yahoo [1], Numenta [2], NASA [3] or Pei's Lab (OMNI) [4], etc. There is a strong implicit assumption that doing well on these public datasets is a sufficient condition to declare an anomaly detection algorithm is useful. In this work, we make a surprising claim. The majority of the individual exemplars in these dataset suffers from one or more of four flaws: *triviality*, *unrealistic anomaly density*, *mislabeled ground truth* and *run-to-failure bias*. Because of these four flaws, we believe that most published comparisons of anomaly detection algorithms may be unreliable, and more importantly, much of the apparent progress in recent years may be illusionary.

II. A Taxonomy of Benchmark Flaws

A. Triviality

If we can quickly create a single line of code (or "one-liner") to separate out anomalies, it strongly suggests that this problem is *trivial*, and that it was not necessary to use several thousands of lines of code and tune a dozen parameters.

To illustrate our point, consider Fig. 1, which shows an example from the OMNI dataset [4]. There are at least three simple one-liners that solve this problem.

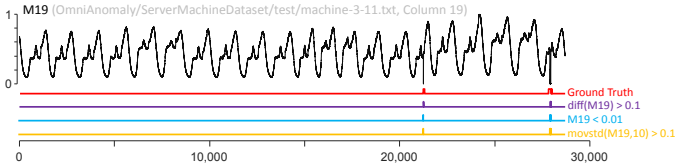

Fig. 1. *(top to bottom)* Dimension 19 from OMNI SDM3-11 dataset. A binary vector (red) showing the ground truth anomaly labels. Three examples of "one-liners" that can solve this problem.

Lest the reader think that we cherry-picked here, let us consider the *entire* Yahoo Benchmark [1], which by far is the most cited in the literature with a mixture of real and synthetic datasets of 367 time series. However, after a simple bruteforce search, we are surprised by its triviality: 316 out of 367 (86.1%) can be easily solved with either (1) or (2).

$$\begin{aligned} abs(\text{diff}(\mathbf{TS})) > \mathbf{u} \times \text{movmean}(abs(\text{diff}(\mathbf{TS}), \mathbf{k}) \\ + \mathbf{c} \times \text{movstd}(abs(\text{diff}(\mathbf{TS})), \mathbf{k}) \\ + \mathbf{b} \end{aligned} \tag{1}$$

$$\begin{aligned} \text{diff}(\mathbf{TS}) > \mathbf{u} \times \text{movmean}(\text{diff}(\mathbf{TS}), \mathbf{k}) \\ + \mathbf{c} \times \text{movstd}(\text{diff}(\mathbf{TS}), \mathbf{k}) \\ + \mathbf{b} \end{aligned} \tag{2}$$

In [5], we show a gallery of dozens of additional examples from Yahoo [1], Numenta [2], NASA [3] and Pei's Lab (OMNI) [4] that yield to one line solutions.

B. Unrealistic Anomaly Density

This issue comes in three flavors:

- More than half the testing data consist of a contiguous region marked as anomalies (NASA D-2, M-1 and M-2).
- *Many* regions are marked as anomalies (OMNI SDM2-5).
- The annotated anomalies are very close to each other (Yahoo A1-Real1).

However, in most real-world settings, the prior probability of an anomaly is expected to be only slightly greater than zero.

We believe that the ideal number of anomalies in a single testing time series is exactly *one*. Instead of trying to predict *if* there is an anomaly in the dataset, the algorithm should just return the most likely *location* of the anomaly.

C. Mislabeled Ground Truth

One of the most referenced datasets is Numenta's NY Taxi data [2]. According to the original labels, there are five anomalies, as annotated in red in Fig. 2.

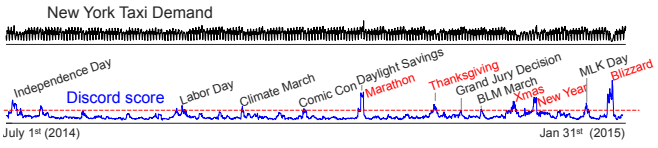

Fig. 2. (*top*) Numenta's NY Taxi dataset. (*bottom*) The *discord* score of the dataset [6], [7], with peaks annotated. The red text denotes the ground truth.

However, these five labels seem very subjective. After a careful visual analysis, we believe that there are at least

seven more events that are equally worthy of being labeled as anomalies, including three additional USA holidays and two marches. The anomaly attributed to the NYC marathon is really caused by a daylight-saving time adjustment.

D. Run-to-failure Bias

There is an additional issue with at least the Yahoo (and NASA) datasets. As shown in Fig. 3, many of the anomalies appear towards the end of the test datasets.

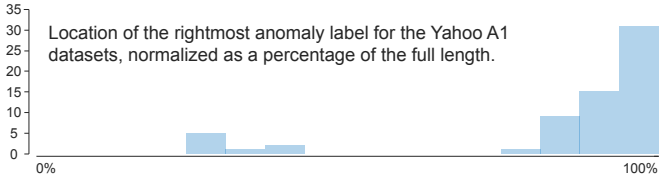

Fig. 3. The locations of the Yahoo A1 anomalies (rightmost, if there are more than one) are clearly not randomly distributed.

A naïve algorithm that simply labels the last point as an anomaly has an excellent chance of being correct.

III. Introducing the UCR Anomaly Archive

Having observed the faults of many existing anomaly detection benchmarks, we have used the lessons learned to create the UCR Time Series Anomaly Archive [8]. As we discussed in Section II-B, we believe that the ideal number of anomalies in a test dataset is one. Below we show two representative examples to explain how we created single anomaly datasets.

A. Natural Anomalies Confirmed Out-of-Band

Consider Fig. 4, which shows an example of one of the datasets in our archive. Here the anomaly is a little subtle.

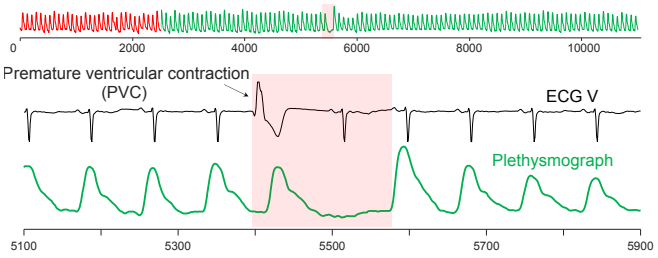

Fig. 4. (*top*) `UCR_Anomaly_BIDMC1_2500_5400_5600`, a dataset from our archive. (*bottom*) A zoom-in of the region containing the anomaly. A PVC observed in the parallel ECG offers out-of-band evidence.

How can we be so confident that it is semantically an anomaly? We can make this assertion because we examined the electrocardiogram that was recorded in parallel. This was the only region that had an abnormal heartbeat, a PVC.

B. Synthetic, but Highly Plausible Anomalies

We can also *create* single anomaly datasets in the following way. We find a dataset that is free of anomalies, then insert an anomaly into a random location. However, we make sure that the resulting dataset is completely plausible and natural.

Fig. 5 shows an example of how we can achieve this. The data came from an individual with an antalgic gait, with a near normal right foot cycle (RFC), but a tentative and weak left foot cycle (LFC). Here we replaced a single, randomly chosen RFC with the corresponding LFC.

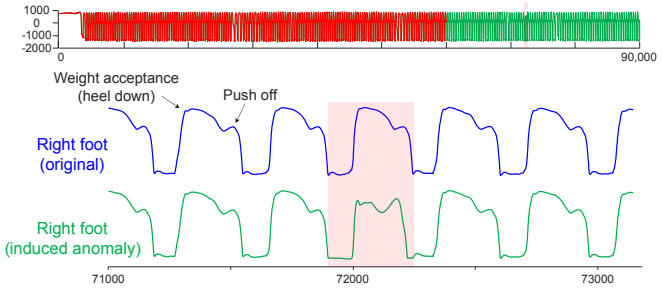

Fig. 5. (*top*) `UCR_Anomaly_park3m_60000_72150_72495`, a dataset from our archive. (*bottom*) This individual had a highly asymmetric gait, so we created an anomaly by swapping in a single *left* foot cycle in a time series that otherwise records the *right* foot.

IV. Conclusions

We have shown that most commonly used benchmarks for anomaly detection have flaws that make them unsuitable for evaluating or comparing anomaly detection algorithms. On a more positive note, we have introduced UCR Time Series Anomaly Archive that is largely free of the current benchmark's flaws [8].

Acknowledgment

The authors wish to thank all the donors of the original datasets, and everyone that provided feedback on this work. We also wish to thank all that offered comments on an early draft of this work, including Matt P. Dziubinski.

References

[1] N. Laptev, S. Amizadeh, and Y. Billawala. (2015) S5 - a labeled anomaly detection dataset, version 1.0 (16M). [Online]. Available: https://webscope.sandbox.yahoo.com/catalog.php?datatype=s&did=70

[2] S. Ahmad, A. Lavin, S. Purdy, and Z. Agha, "Unsupervised real-time anomaly detection for streaming data," *Neurocomputing*, vol. 262, pp. 134–147, 2017.

[3] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, and T. Soderstrom, "Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding," in *Proc. 24th ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining*, 2018, pp. 387–395.

[4] Y. Su, Y. Zhao, C. Niu, R. Liu, W. Sun, and D. Pei, "Robust anomaly detection for multivariate time series through stochastic recurrent neural network," in *Proc. 25th ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining*, 2019, pp. 2828–2837.

[5] R. Wu and E. J. Keogh. (2021) Supporting page for current time series anomaly detection benchmarks are flawed and are creating the illusion of progress. [Online]. Available: https://wu.renjie.im/research/anomaly-benchmarks-are-flawed/

[6] T. Nakamura, M. Imamura, R. Mercer, and E. Keogh, "MERLIN: Parameter-free discovery of arbitrary length anomalies in massive time series archives," in *Proc. 2020 IEEE Intl. Conf. Data Mining*, 2020, pp. 1190–1195.

[7] D. Yankov, E. Keogh, and U. Rebbapragada, "Disk aware discord discovery: Finding unusual time series in terabyte sized datasets," in *Proc. 7th IEEE Intl. Conf. Data Mining*, 2007, pp. 381–390.

[8] E. Keogh, T. D. Roy, U. Naik, and A. Agrawal. (2021) Multi-dataset time series anomaly detection competition, SIGKDD 2021. [Online]. Available: https://compete.hexagon-ml.com/practice/competition/39/