Detecting Anomalies in Spacecraft Telemetry Using Evolutionary Thresholding and LSTMs

Pawel Benecki
KP Labs/Silesian University of Technology
Gliwice, Poland
pbenecki@kplabs.pl

Szymon Piechaczek
KP Labs
Gliwice, Poland
spiechaczek@kplabs.pl

Daniel Kostrzewa
KP Labs/Silesian University of Technology
Gliwice, Poland
dkostrzewa@kplabs.pl

Jakub Nalepa
KP Labs/Silesian University of Technology
Gliwice, Poland
jnalepa@ieee.org

ABSTRACT

Detecting anomalies in telemetry data captured on-board satellites is a pivotal step towards their safe operation. The data-driven algorithms for this task are often heavily parameterized, and the incorrect hyperparameters can deteriorate their performance. We tackle this issue and introduce a genetic algorithm for evolving a dynamic thresholding approach that follows a long short-term memory network in an unsupervised anomaly detection system. Our experiments show that the genetic algorithm improves the abilities of a detector operating on multi-channel satellite telemetry.

CCS CONCEPTS

• **Computing methodologies** → **Genetic algorithms**; *Anomaly detection*; *Neural networks*; • **Applied computing** → *Forecasting*;

KEYWORDS

Genetic algorithm, anomaly detection, satellite telemetry, LSTM

ACM Reference Format:
Pawel Benecki, Szymon Piechaczek, Daniel Kostrzewa, and Jakub Nalepa. 2021. Detecting Anomalies in Spacecraft Telemetry Using Evolutionary Thresholding and LSTMs. In *2021 Genetic and Evolutionary Computation Conference Companion (GECCO '21 Companion), July 10–14, 2021, Lille, France.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3449726.3459411

1 INTRODUCTION

Detecting anomalies in spacecraft telemetry is a critical aspect of its safe operation. There are three main types of anomalies that should be considered for complex missions—in *point* anomalies, telemetry values fall outside the nominal operational range. The *collective* anomalies refer to the overall sequences of consecutive telemetry values that are anomalous, whereas in *contextual* anomalies, the single values are anomalous within their neighborhood [1].

GECCO '21 Companion, July 10–14, 2021, Lille, France

ACM ISBN 978-1-4503-8351-6/21/07…$15.00
https://doi.org/10.1145/3449726.3459411

The out-of-limit detection engines can spot point anomalies, and various expert systems cover other events [2]. Since generating new ground-truth sets is extremely costly, unsupervised algorithms have become the mainstream. In Telemanom [1], the expected telemetry values are extracted using long short-term memory (LSTM) nets. Then, an unsupervised thresholding of differences between the expected and actual values is used to detect events. As separate LSTMs process different telemetry channels, Telemanom offers traceability and interpretability which are crucial in Space applications.

Data-driven algorithms for detecting telemetry anomalies are commonly heavily parameterized, and the incorrect hyperparameters deteriorate their performance. We build upon [1], and propose a genetic algorithm (GA) to evolve the hyperparameters of its unsupervised thresholding part (Sect. 2). The experiments indicate that GA improves the Telemanom's abilities (Sect. 3). We show that assessing the quality of detectors should be revisited, as the metrics that capture temporal aspects of detected anomalies (with respect to the ground truth) convey very important information.

2 GENETIC OPTIMIZATION OF DYNAMIC THRESHOLDING PARAMETERS

In our GA, an initial population of N individuals p_i, where $i = 1, 2, \ldots, N$, each encoding a set of hyperparameters (Table 1; for details, see [1]), is evolved. The values for each p_i are randomly sampled from the uniform distribution for each hyperparameter, according to their pre-defined feasible ranges. The **fitness** η quantifies the quality of the underlying parameterization of the thresholding calculated over the set of training sequences $\boldsymbol{T}$. To ensure safe operation of a spacecraft, the anomalies should be detected as fast as possible to timely take actions, and the number of false positives should be low, especially in semi-automatic systems, in which the action is taken by a human. We exploit DICE and F_β as the fitness, and $\mathrm{DICE}(A, B) = 2 \cdot |A \cap B| / (|A| + |B|)$, where A and B are two anomalous regions, i.e., manual and automated, within the signal, whereas $F_\beta = \left[(1+\beta^2) \cdot \mathrm{TP}\right] / \left[(1+\beta^2) \cdot \mathrm{TP} + \beta^2 \cdot \mathrm{FN} + \mathrm{FP}\right]$, where TP, FN, and FP are true positives, false negatives, and false positives in the anomaly. To prioritize precision, we have $\beta = 0.5$.

In **crossover**, we generate two children for each (out of $N/2$) pair of parental solutions (p_a, p_b), where $a \neq b$, and each chromosome is selected to act as a parent from $(C_s \cdot N)$ best individuals. Here, we always pre-select p_{best} to be included in at least one parental pair.

The hyperparameter values in offspring individuals are inherited with an equal probability from each parent. Next, the first child is **mutated** through changing the hyperparameters (with probability $\mathcal{P}_m$) to the random ones drawn from the corresponding range of feasible values. In the second child, we pick a random hyperparameter from the range limited by the parents' current values. All children, alongside p_{best} survive to the next generation. The evolution continues until the maximum number of generations $G_{\max}$ have been processed, or $\eta(p_{\text{best}})$ has not changed for G_{ES} generations.

Table 1: The hyperparameters that undergo evolution. We include the default and GA-evolved (for best DICE) values.

Symbol	Range	Step	Def.	GA	Meaning
n_{B}	$[10, 200]$	10	70	110	Number of values analyzed in a single telemetry batch.
n_{W}	$[5, 100]$	5	30	50	The number of consecutive batches analyzed together.
n_{E}	$[0.005, 0.1]$	Cont.	0.05	0.01	The window size used in the error smoothing (being the percentage of all values in a series).
$\boldsymbol{e}$	$[50, 500]$	10	100	50	The number of values surrounding errors (it may promote grouping of nearby error sequences).
p	$[0.05, 0.2]$	Cont.	0.13	0.06	Min. percent difference between subsequent anomalies.

3 EXPERIMENTS

We exploit the set of 82 telemetry time series, where each sequence is split into train (without anomalies) and test (containing manually annotated anomalies, with their starting and termination points) parts [1]. The test sequences are of variable length ($410 - 8380$, with an average of 5962). GA was run 20× for each configuration (DICE and $F_{0.5}$ as η), and its hyperparameters were manually-tuned and kept unchanged through the experimentation: $N = 20$, $G_{\max} = 10^3$, $G_{\text{ES}} = 10$, $\mathcal{P}_m = 0.1$, and $C_s = 0.3$ (we executed GA with and without early stopping, the G_{ES} and $G_{\max}$ variants, respectively). Every fifth sequence is included in $\boldsymbol{T}$ which is used for quantifying the fitness. We report the results averaged over all sequences.

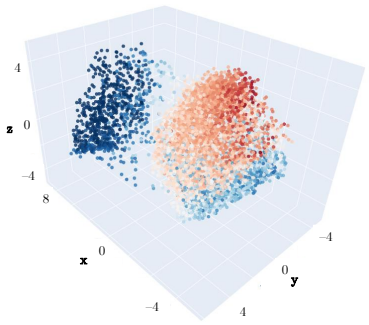

Figure 1: There are hyperparameters that render significantly better DICE. We show three Isomap dimensions [3]. Dark blue shows the highest DICE, dark red—the lowest.

We confronted GA with the default hyperparameters [1], and with random search (RS) in which we perform $N \cdot G_{\max} = 2 \cdot 10^4$ evaluations. There are parts of the solution space encompassing hyperparameters that render notably better scores (Figure 1).

In Table 2, we gather DICE and $F_{0.5}$ averaged across all series. Also, we present the average ($\Delta_{\text{B}}^{\text{avg}}$) and maximum ($\Delta_{\text{B}}^{\max}$) difference between the starting points of the detected and ground-truth anomalies. To increase the usability of the detector in practical spacecraft scenarios, both $\Delta_{\text{B}}^{\text{avg}}$ and $\Delta_{\text{B}}^{\max}$ should be minimized—their negative values indicate that the sequence was annotated as anomalous before the anomaly actually happened (in such cases, we *predict* anomalies). The results show that evolved hyperparameters elaborate the high-quality detection. Also, GA converges much faster than RS—with our early stopping, we executed 28× and 40× less evaluations than RS for DICE and $F_{0.5}$ as η, respectively.

Table 2: The results obtained over all sequences (we report the average DICE and $F_{0.5}$). For each metric, the best value is boldfaced, and the second best is underlined.

Algorithm	DICE	$F_{0.5}$	$\Delta_{\text{B}}^{\text{avg}}$	$\Delta_{\text{B}}^{\max}$	Evaluations
Default	0.3945	<u>0.8071</u>	−78	141	–
Random search	<u>0.4639</u>	**0.8535**	**−205**	433	$2.00 \cdot 10^4$
GA(DICE, G_{ES})	0.4289	0.7686	<u>−110</u>	**20**	$\mathbf{0.05 \cdot 10^4}$
GA(DICE, $G_{\max}$)	**0.4659**	0.8052	−75	<u>77</u>	$2.00 \cdot 10^4$
GA($F_{0.5}$, G_{ES})	0.3763	0.7566	−98	334	<u>$0.07 \cdot 10^4$</u>
GA($F_{0.5}$, $G_{\max}$)	0.3641	0.7592	−81	121	$2.00 \cdot 10^4$

Although RS delivered large $F_{0.5}$, there exist anomalous sequences in which the events were spotted at their late stage (see $\Delta_{\text{B}}^{\max}$)—it hampers the applicability of this approach in practice. Our GA not only retrieved the parameterizations that enable us to capture the anomalies with a significant time margin on average, but also the GA variants with the DICE fitness led us to the best $\Delta_{\text{B}}^{\max}$'s. Finally, the experiments show that introducing new temporal-based metrics to assess the quality of anomaly detection is pivotal in understanding its applicability, as large DICE or $F_{0.5}$ would not reflect how fast an anomaly can be detected (or even predicted), hence it is unclear how much time we can allow for taking appropriate actions.

4 CONCLUSIONS

We introduced a GA to evolve the hyperparameters of a thresholding step in a deep learning anomaly detection engine, and showed that it improves the abilities of this technique. We shed more light on the problem of assessing the anomaly detectors, and we claim that incorporating the temporal aspect in the quantitative metrics is crucial in Space applications.

ACKNOWLEDGMENTS

This work was supported by the Polish National Centre for Research and Development (POIR.01.01.01-00-0853/19), and by the Silesian University of Technology grant for maintaining and developing research potential (BKM21) and Rector's grant (02/080/RGJ20/0003).

REFERENCES

[1] Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. 2018. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *Proc. ACM SIGKDD*. 387–395.
[2] Haixu Jiang, Ke Zhang, Jingyu Wang, Xianyu Wang, and Pengfei Huang. 2020. Anomaly Detection and Identification in Satellite Telemetry Data Based on Pseudo-Period. *Applied Sciences* 10, 1 (2020). https://doi.org/10.3390/app10010103
[3] Yan Zhang, Zhao Zhang, Jie Qin, Li Zhang, Bing Li, and Fanzhang Li. 2018. Semi-Supervised Local Multi-Manifold Isomap by Linear Embedding for Feature Extraction. *Pattern Recognition* 76, C (2018), 662–678.