17th International Conference on Space Operations, Dubai, United Arab Emirates, 6 - 10 March 2023.

SpaceOps-2023, ID # 169

Compact Distributions With Self-ensembled Scoring for Satellite Anomaly Detection

Guo Guohang[a,b]*, Hu Tai[a], Liu Yurong[a], Yin Xiaodan[a,b], Guo Yuting[c], Li Hu[a]

[a] *National Space Science Center, Chinese Academy of Sciences, Beijing 100190, China*
[b] *University of Chinese Academy of Sciences, Beijing 100049, China*
[c] *Ocean University of China, Shandong 266100, China*
* Corresponding Author

Abstract

Satellites are exceptionally complex and expensive machines with thousands of telemetry channels. Monitoring these channels is an important and necessary component of satellite operations given the complexity and cost of the satellite. As satellites send back increasing amounts of telemetry data, improved anomaly detection methods are needed to lessen the monitoring burden placed on operations engineers and reduce operational risk. Contrastive learning-based anomaly detection method is widely explored in nowadays research. While contrastive learning-based anomaly detection methods faced the following problems in satellite telemetry data anomaly detection scenario: 1) Telemetry data has high dimensions and may contain irrelevant features, where these noisy dimensions could hide the characteristics of anomaly data and thus make the detection process more difficult. 2) The feature distribution learned by contrastive learning is indistinguishable and hard to discriminate normal and anomalies in feature space. 3) The construction of the anomaly score only considers the single-view features in the proxy task of contrastive learning, which loses the anomaly decision-making information. To solve the above problems, we propose a method named compact distributions with self-ensembled scoring for satellite anomaly detection. We first adopt LightGBM to select features related to the status of the satellite according to the feature importance ranking. Then we combine the contrastive loss and the center loss to obtain a more compact feature distribution, in which the normal and abnormal samples are easier to distinguish. Finally, we build probability distribution of each view's feature using KDE(kernel density estimation) and thus construct anomaly score. We conduct experiments on telemetry data and our method achieves state-of-the-art anomaly detection performance compared with multiple baseline methods.
Keywords: Telemetry data; Anomaly detection; Contrastive learning

1. Introduction

Satellites are exceptionally complex and expensive machines with thousands of telemetry channels detailing aspects such as temperature, radiation, power, instrumentation, and computational activities[1-3]. Monitoring these channels is an important and necessary component of satellite operations given the complexity and cost of the satellite. As satellites send back increasing amounts of telemetry data, improved anomaly detection methods are needed to lessen the monitoring burden placed on operations engineers and reduce operational risk.

Anomaly detection, a.k.a. outlier detection or novelty detection, is referred to as the process of identifying patterns that do not conform to the expected normal patterns[4, 5]. Anomaly detection is widely applied in network intrusion detection, fraud detection, industrial process monitoring, surveillance videos, and numerous other fields. Anomaly detection can raise an alarm to remind people to pay more attention when a different pattern occurred which never seen before. This is the fundamental difference from the supervised learning which can get all data classes.

There are different possible scenarios for anomaly detection. In a supervised setting, we are given training samples of anomaly and normal patterns. In this case, the problem simplifies to the imbalanced version of data classification. However, obtaining such supervision may not be possible all the time. For example in a spacecraft anomaly detecting setting, some anomalies may be known, but other new types of anomalies are continually discovered over time owing to changing environmental factors and command sequences. On the other extreme, fully unsupervised anomaly detection obtains training samples with normal and anomaly patterns and attempts to detect anomalous data. In this work, we deal with the semi-supervised scenario, in which only samples of the normal pattern are available during the training phase. After training the anomaly detector, we detect anomalies in the test data which contain both normal and anomaly samples. Such a setting is quite common and easy to obtain in practical scenarios.

Many anomaly detection literatures have been proposed in the past decades, including reconstruction-based, density-based, one-class classification, and self-supervised methods. The majority of recent research centers on (a)

learning high-level data representations to encode normal samples, and (b) learning a detection score based on the representation. Following the success of deep learning, a range of deep anomaly detection approaches have been proposed, which rely on the power of deep methods to get a compact low-dimensional representation around a center on the training samples. However, normal and anomaly samples in the test set are mapped in an overlapped latent space, which leads to indiscriminate features. Previous works circumvent this issue by adopting variant autoencoder pretraining, adversarial training, and adversarial samples injecting.

Meanwhile, recent progress on self-supervised learning has proven the effectiveness of contrastive learning in computer vision and audio processing. Contrastive learning aims at learning a feature representation such that multiple views(e.g. augmentations) of a sample keep attracting while repelling to other samples. Contrasting shifted instances, a special type of contrastive learning, has adapted the existing contrastive learning schemes to anomaly detection settings and achieved state-of-the-art results.

In this work, we present a three-stage framework: (1) adopt LightGBM to select features related to the status of the satellite according to the feature importance ranking, (2) train a deep network to obtain a (more) discriminative representation and (3) build probability distribution of each view's feature using KDE and thus construct anomaly score.

2. Proposed Method

In the semi-supervised scenario, we are given a training dataset $\mathcal{X}_{train}$ with just normal samples. The goal of anomaly detection is to model a detector from $\mathcal{X}_{train}$ that determines whether a new sample x is normal or not. The method aims to obtain a deep representation of a data sample by the neural network function $\phi : \mathcal{X} \to \mathbb{R}^d$, where d is the feature dimension. In this work, ϕ is learned by using a contrastive learning task on $\mathcal{X}_{train}$. Then, an anomaly scoring function $s(\phi(x))$ predicts how anomalous the sample x is. In this section, we propose a two-stage approach for anomaly detection. In section 2.1, we present our new loss function which contains masked supervised contrastive loss and cosine center loss. In section 2.2, we present an ensemble anomaly scoring function by using KDE.

2.1 Background: Supervised Contrastive Learning

The mainline of the top-performing methods is to learn the representations based on contrastive learning, especially simple contrastive learning(SimCLR)[6]. Supervised contrastive learning(SupCLR) [7] is a supervised variant of SimCLR that contrasts samples class-wise rather than instance-wise: any sample pair of the same classes are deemed positive.

The main components of SupCLR are data augmentation module $\mathcal{A}(\cdot)$, feature extractor $f(\cdot)$, and projection head $g(\cdot)$. $\mathcal{A}(\cdot)$ is a random data augmentation process, which generates augmentation $x' = \mathcal{A}(x)$ of input sample x. $f(\cdot)$ parameterized by deep neural networks maps x to a representation $r = f(x)$, which is used for the downstream task. $g(\cdot)$ maps r to a vector $z = g(r)$, which is used to compute proxy loss $\mathcal{L}_{supclr}$ so that $f(\cdot)$ outputs task-specific representations.

In the training procedure, a batch of sample pairs with size N, $\{(x_k, y_k)\}_{k=1...N}$, is randomly sampled and the contrastive prediction task is defined on $2N$ pairs, $\{(\tilde{x}_m, \tilde{y}_m)\}_{m=1...2N}$, where $\tilde{x}_{2k} = \mathcal{A}_1(x_k)$ and $\tilde{x}_{2k-1} = \mathcal{A}_2(x_k)$ are two different random augmentations(a.k.a., "views") of $x_k (k = 1...N)$ and $\tilde{y}_{2k} = \tilde{y}_{2k-1} = y_k$. Let $i \in I \equiv \{1...2N\}$ be the index of a sample in the augmented batch, and $P(i) \equiv \{p \in I \setminus \{i\} : \tilde{y}_p = \tilde{y}_i\}$ be the set of indices of all positive augmented samples distinct from i. Following [7], the proxy task loss of SupCLR is written as:

$$\mathcal{L}_{supclr} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \frac{exp(sim(z_i, z_p)/\tau)}{\sum_{k=1}^{2N} \mathbb{I}_{[k \neq i]} exp(sim(z_i, z_k)/\tau)} \tag{1.1}$$

where $\mathbb{I}_{[k \neq i]} \in \{0,1\}$ is an indicator function evaluating to 1 iff $k \neq i$, τ denotes a positive temperature parameter, and $sim(z_i, z_p) = z_i^T z_p / (\| z_i \| \| z_p \|)$ denotes the dot production between l_2 normalized z_i and z_p (i.e. cosine similarity).

Contrastive methods have recently achieved superior performance in a variety of fields, especially visual recognition tasks. While we argue that it could be problematic for one-class classification which needs a discriminative representation with low intra-class variation and high inter-class variation. First, the denominator of Eq.1 is minimized by repelling the representations of sample pairs, even though they are from the same class. This seems to contradict the idea of deep one-class classification. Second, the loss of Eq.1 has no mechanism to constrain the distance of samples from its class center. While in the one-class classification setting, we expect the normal data to lie in a small region around the center which will be more discriminative than it is allowed to occupy a larger distance.

2.2 The Mean-Shifted Contrastive Loss

Different from the setting of the vanilla SupCLR which regards different augmentations as the same class, we follow the setting of instance discrimination in [8] that considers different augmentations to have different class labels and the same augmentation has the same class label. That means in this work, we have the following setting: in the augmentation batch, $\tilde{x}_{2k} = \mathcal{A}_1(x_k)$, $\tilde{x}_{2k-1} = \mathcal{A}_2(x_k)$, and $\tilde{x}_{2l} = \mathcal{A}_1(x_l)$, $\tilde{x}_{2l-1} = \mathcal{A}_2(x_l)$ are respectively two random augmentations of x_k and x_l, where $k,l = 1...N$ and $k \neq l$, we deem $y_{2k} = y_{2l} = C_1$ and $y_{2k-1} = y_{2l-1} = C_2$ where C_m denotes the class label of the augmentation $\mathcal{A}_m(x)$.

To solve the first problem mentioned above, we add a component named self-adaption temperature($\mathcal{SAT}$) to adaptively determine the repelling ratio considering the label information. $\mathcal{SAT}$ can be defined as follows:

$$\mathcal{SAT}(i,j) = \begin{cases} \tau & if\ \tilde{y}_i \neq \tilde{y}_j \\ \alpha & if\ \tilde{y}_i = \tilde{y}_j \end{cases} \tag{1.2}$$

where $\tau < \alpha$. $\mathcal{SAT}$ raises the temperature for views with the same label to a higher value α, causing the query view to repel views with the same label by a little amount when compared to views with different labels. The modified loss function of SupCLR is as follows:

$$\mathcal{L}_{sat} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \frac{exp(sim(z_i, z_p)/\tau)}{\sum_{k=1}^{2N} \mathbb{I}_{[k \neq i]} exp(sim(z_i, z_k)/\mathcal{SAT}(i,k))} \tag{1.3}$$

For the sake of expression, we define repelling ratio as $1/\mathcal{SAT}(i,j)$. The loss function penalizes the views with the same class label with a large temperature resulting in a small repelling ratio, which will obtain a lower inter-variation than that of the different views. The reason why we penalize the views in the same class with a small repelling ratio rather than 0 is the repelling ratio of 0 has the risk of incurring hyper collapse like the issue in DeepSVDD [9].

To obtain a compact representation for the downstream task, we define the center loss. Different from previous works, we define the center loss on the $f(x)$ rather than normalized $g(x)$, and the experiment in section 3 demonstrates this. The center loss is defined as follows:

$$\mathcal{L}_{center} = \sum_{m=1}^{|C|} \sum_{i=1}^{N} \| f(\mathcal{A}_m(x_i)) - c_m \|^2 \tag{1.4}$$

where C is the set of class label, $|C|$ is the number of classes(a.k.a. augmentations), and the center c_m are given by the average feature over the training set for every transformation i.e. $c_m = \frac{1}{N} \sum_{i=1}^{N} f(\mathcal{A}_m(x_i))$.

The final loss of our proposed method is defined by combining the two losses:

$$\mathcal{L}_{joint} = \mathcal{L}_{sat} + \lambda \mathcal{L}_{center} \tag{1.5}$$

where positive λ is a hyper-parameter weighting the two losses. We set $\lambda = 5$ in all our experiments.

2.3 Score functions for detecting anomalous

We use the learned low-representation $f(x)$ to model a probability density distribution of input data. We further utilize the features to calculate their probability density distribution through KDE model.

Following the kernel density estimation model, the probability density distribution function of data is as follows:

$$f_h(s) = \frac{1}{n}\sum_{i=1}^{n} K_h(s - f(x_i)) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{s - f(x_i)}{h}) \tag{1.6}$$

Where s is the variable, and K is the kernel(a non-negative function) and $h(h > 0)$ is a smoothing parameter called the bandwidth. In this paper, we adopt Gaussian kernel function in the KDE model.

3. Experiments

In this section, we employ the actual telemetry data from the Quantum Science Experiment Satellite a.k.a. Micius to demonstrate the effectiveness of our method. The dataset contains 19 attributes related to the operation pattern of the satellite, and the time span is from January 2017 to February 2019. Micius has 4 operation patterns. As pattern 4 is rare, we treat pattern 4 as anomaly class and the rest as normal class.

3.1 Baseline Methods

OC-SVM [10]. OC-SVM(One-Class Support Vector Machine) is a popular kernel method used in anomaly detection, which aims to learn a decision boundary only using the normal data. We adopt the widely used radial basis function(RBF) in this work.

LOF [11]. LOF(Local Outlier Factor) is an unsupervised anomaly detection method. The anomaly score depends on how isolated the object is with respect to the surrounding neighborhoods.

DAGMM [12]. DAGMM(Deep Autoencoding Gaussian Mixture Model) is a widely used method for anomaly detection, which contains a compression network and an estimation network. The compression network utilizes a deep autoencoder to generate a low-dimensional representation and the representation is fed into the estimation network to obtain an energy score.

Deep SVDD [9]. Deep SVDD(Deep Support Vector Data Description) is the deep variant of SVDD, which aims to leverage the power of deep learning to learn a hypersphere only using normal data.

GOAD[13]. GOAD is a classification-based method for detecting anomalies for general data, which obtains anomaly scores by training a classifier on a set of random auxiliary tasks.

3.2 Evaluation metrics

Following [12] and [13], the methods are trained on a random subset of 50% of the normal data and are evaluated on the remaining normal data as well as all the anomaly data.

We use the mean and standard deviation(σ) of F_1 score after 20 random splits to compare the anomaly detection performance. We take anomaly class as positive, and define F_1 score accordingly. F_1 score is defined as follows: $F_1 = \frac{2 * Precision * Recall}{Precision + Recall}$, where Precision= $\frac{|G \cap R|}{|R|}$, Recall= $\frac{|G \cap R|}{|G|}$, G denotes the set of ground truth anomalies in the dataset, and R denotes the set of anomalies reported by the methods.

3.3 Effectiveness Evaluation

We evaluate the effectiveness of our proposed method compared with 5 baseline methods.

Table 1. F_1 score and standard deviation of our method and baselines on Micius

	OC-SVM	LOF	DAGMM	Deep SVDD	GOAD	Ours
$F_1 \pm \sigma$	53.83±0.15	47.80±4.54	49.61±7.10	16.46±9.99	39.91±5.55	60.46±3.59

Table1 shows the average F_1 score with standard deviation for all methods. Our method outperforms the literature baselines by a significant margin, which achieves 12.3% improvement compared to the state-of-the-art OC-SVM.

To our surprise, the two classical baseline methods OC-SVM and LOF achieve good performance on Micius even surpassing some deep methods. The classical method OC-SVM achieves the next best performance on Micius out of all the methods. This is because when extracting features, the majority of deep approaches do not emphasize maintaining the local structure of the data. It is difficult to detect local anomalies in such a dataset, while OC-SVM detects anomalies by considering the local density information of the data. Fortunately, our method takes into account the local information through the prototype information of the data, which enables it to accurately detect anomalies on the Micius dataset. Therefore, data characteristics and the type of anomalies should be taken into account in anomaly detection tasks.

4. Conclusions

In this paper, We proposed a method to detect anomalies in telemetry data. We first adopt LightGBM to select features related to the status of the satellite. Then we combine the contrastive loss and the center loss to obtain a more compact to distinguish the normal and abnormal samples. Finally, we build probability distribution of each view's feature using KDE(kernel density estimation) and thus construct anomaly score. We conduct experiments on telemetry data and our method achieves state-of-the-art anomaly detection performance compared with multiple baseline methods.

References

[1] Baireddy, S., et al. *Spacecraft time-series anomaly detection using transfer learning*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.
[2] Hassanien, A.E., A. Darwish, and S. Abdelghafar, *Machine learning in telemetry data mining of space mission: basics, challenging and future directions.* Artificial Intelligence Review, 2020. **53**(5): p. 3201-3230.
[3] Ji, X.-Y., et al., *A brief review of ground and flight failures of Chinese spacecraft.* Progress in Aerospace Sciences, 2019. **107**: p. 19-29.
[4] Pang, G., et al., *Deep learning for anomaly detection: A review.* ACM Computing Surveys (CSUR), 2021. **54**(2): p. 1-38.
[5] Zhang, X., et al., *Deep anomaly detection with self-supervised learning and adversarial training.* Pattern Recognition, 2022. **121**: p. 108234.
[6] Chen, T., et al. *A simple framework for contrastive learning of visual representations*. in *International conference on machine learning*. 2020.
[7] Khosla, P., et al., *Supervised contrastive learning.* Advances in Neural Information Processing Systems, 2020. **33**: p. 18661-18673.
[8] Tack, J., et al., *Csi: Novelty detection via contrastive learning on distributionally shifted instances.* Advances in neural information processing systems, 2020. **33**: p. 11839-11852.
[9] Ruff, L., et al. *Deep one-class classification*. in *International conference on machine learning*. 2018.
[10] Chen, Y., X.S. Zhou, and T.S. Huang. *One-class SVM for learning in image retrieval*. in *Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)*. 2001.
[11] Breunig, M.M., et al., *LOF: Identifying Density-Based Local Outliers.* SIGMOD Rec., 2000. **29**(2): p. 93–104.
[12] Zong, B., et al. *Deep autoencoding gaussian mixture model for unsupervised anomaly detection*. in *International conference on learning representations*. 2018.
[13] Bergman, L. and Y. Hoshen, *Classification-based anomaly detection for general data.* arXiv preprint arXiv:2005.02359, 2020.