List of Changes and Answers to Reviewers

Adaptable and Interpretable Framework for Anomaly Detection in SCADA-based Industrial Systems
M. Wadinger, M. Kvasnica

November 28, 2023

1 List of Changes

List of main changes:

1. Figure 2 was added to illustrate the definition of anomalies to address reviewers' comments.
2. Captions in Figures 5,6,7 (labeled as Figures 4,5,6 in the previous manuscript) were reworded to make them more illuminating.
3. Added other self-supervised change-point detection method from [DSXS21] in the Introduction.
4. Table 2 was transposed to allow for new metrics to be added.
5. False Alarm Rate, AUC, and Mean of Rolling AUC were added to the results in Table 2 to address reviewers' comments.
6. Section 4.4 was added to address reviewers' suggestion on scalability analysis.

2 Answers to Reviewers and to the Associate Editor

We would like to thank all reviewers and to the associate editor for encouraging comments and hints. We have tried to address all of them appropriately.

Associate Editor

Reviewers have now commented on your paper. You will see that they are advising that you revise your manuscript. If you are prepared to undertake the work required, I would be pleased to reconsider my decision.

For your guidance, reviewers' comments are appended below.

If you decide to revise the work, please submit a list of changes or a rebuttal against each point raised by the reviewers. You can upload this as the 'Detailed Response to Reviewers' when you submit the revised manuscript.

Response: We would like to thank the associate editor for his/her evaluation. We believe that the modifications, described in more detail below, address all issues pointed out by the reviewers.

Reviewer 1

This paper proposes a new online anomlay detection method and verifies its effectiveness on real-world datasets. However, there are some limitations as follows:

1. *There is no clear definitions of point anomaly, collective anomalies and concept changes. Figure 2 does not illustrate their differences either.*

   Response: While we tried to establish the notation of anomalies and their categorization used throughout our paper in Section 2.6, the visualization would be more illuminating.

   Indeed, the work would benefit from more clarity in defining point anomalies, collective anomalies, and concept changes.

   Therefore, we have added Figure 2 in the revision to illustrate the distinctions between these anomaly types. Figure 2, previously in question, is now labeled as Figure 3 in the revision.

2. *The captions of Figures 4,5,6 are similar but the labels are different. It is a bit confused as which one is the ground truth.*

   Response: We appreciate the reviewer's keen observation regarding the similarity in captions for Figures 4, 5, and 6. Recognizing the importance of conveying clear messages through captions, we have revised the captions to eliminate any confusion. Please note that Figures 4, 5, and 6 in the initial submission are now labeled as Figures 5, 6, and 7 in the revised manuscript due to the addition of Figure 2 to address the previous reviewer's comment.

   The changes made to the captions include a description of the experiment's setup and the proposed method's allowed features to illuminate their contribution to the whole framework.

   It is important to note that the challenge of obtaining precise ground truth information remains. Operators did not inform us about the exact time of abnormal events, introducing ambiguity. While we refer to the dates of

the events in Section 4.1, selecting the time of the anomaly for plotting purposes would be arbitrary and compromise the objectivity of the results.

3. *There are other self-supervised change-point detection method, such as [DSXS21].*

   Response: We appreciate the reviewer's suggestion and acknowledgment of the reference [DSXS21]. Our primary goal was to offer a comprehensive review of state-of-the-art self-supervised adaptive detection methods with interpretability. While our review aimed to cover the most relevant self-supervised change-point detection methods, we acknowledge that it may not be exhaustive in this regard.

   Upon careful examination of the suggested reference, we found it to be highly relevant, particularly in the paragraph discussing the need for an early change point detection mechanism. We have incorporated the reference into the Introduction section of the paper to enrich the discussion.

4. *It seems that using ARIMA or moving average can easily detect the anomalies or change points on the real-world datasets.*

   Response: While we acknowledge that the figures may suggest that anomalies are easily detected by ARIMA or moving average methods, we want to emphasize the unique features of our proposed method that differentiate it in the context of online anomaly detection for evolving data streams.

   Our proposed method considers interactions between variables, providing diagnostic capabilities that may be crucial in real-world scenarios. As part of our extensive literature review, we discovered that Vector Autoregression, the multivariate extension of ARIMA, shows promise for anomaly detection in multivariate time series by capturing complex interactions. Notably, papers such as [MBMO16, ZZQ23] explore offline trained anomaly detection methods based on vector autoregression.

   Our research focuses on online anomaly detection for evolving data streams, a problem requiring a unique combination of features relevant to real-world scenarios. Although we are aware of the existence of online-trained ARIMA methods [AHMS13], a vectorized implementation of online-trained ARIMA is currently lacking. This limitation impeded its inclusion in our comparison.

   We believe that integrating our proposed method with ARIMA during feature engineering, as shown in Section 4.1 for physics-based model utilization, showcases a promising direction for enhancing performance. Future research could explore extending ARIMA to its multivariate counterpart for online training.

5. *It would be better to use AUC rather than F1 to measure the anomaly detection performance. Moreover, range-based AUC is even better and more fair for streaming or sliding window-based method.*

   Response: We agree with the reviewer that AUC, in general, is a better metric for imbalanced datasets. However, due to the poor convergence of the reference methods on benchmark data during hyperparameter optimization with AUC, we decided to use the F1 score, which showed better convergence for all three compared methods.

   In response to the reviewer's recommendation, we have included AUC in the results presented in Table 2 of the revised paper. The addition of AUC provides an alternative perspective on performance that may be of interest to the reader.

   Additionally, we were not aware of the range-based AUC metric during the time of paper writing and result collection. After discovering this suggestion, we computed the mean value of range-based AUC using the implementation from [BS17] and enriched the results in Table 2.

   We also attempted to use the mean value of range-based AUC for hyperparameter optimization. However, due to minimal improvement in convergence compared to regular AUC, we decided to retain the F1 score as the optimized metric. The results obtained using the range-based AUC metric in the hyperparameter optimization cost function are provided in Table 1 for reference.

   We hope these additions and explanations enhance the transparency and completeness of our evaluation.

Table 1: Evaluation of models optimized for Rolling AUC score on SKAB dataset. The best-performing model is highlighted in bold. Values in brackets represent macro values of the metric.

Algorithm	AID	HS-Trees	OC-SVM
Precision [%]	**47** (60)	30 (47)	32 (48)
Recall [%]	**55** (61)	4 (50)	3 (50)
F1 [%]	**51** (60)	7 (42)	6 (42)
AUC [%]	**61**	50	50
Mean Rolling AUC [%]	**60**	50	49
FPR [%]	38	**37**	**37**

Reviewer 2

This paper presents an interesting and potentially useful framework called AID for anomaly detection and root cause diagnosis in industrial internet-of-things (IoT) systems. It incorporates dynamic conditional probability distribution modeling to adapt to non-stationary data streams, which is crucial for industrial systems. And industrial case studies demonstrate capabilities on real systems. However, i still have following concerns.

1. *More analysis of computational complexity and scalability limitations for high-dimensional industrial systems would strengthen the work.*

   Response: Thank you for bringing attention to the importance of computational complexity and scalability in the context of high-dimensional industrial systems. To address this suggestion, we have incorporated Section 4.4 into our paper, providing a dedicated analysis of scalability and time complexity.

   In Section 4.4, we delve into the computational complexities and scalability limitations of our proposed method, in the context of high-dimensional industrial systems. The section aims to offer insight into the scalability of the proposed method for anomaly detection tasks and in the context of SCADA-based systems, where dynamic limits are of interest. We hope that this addition will enhance the completeness of our work.

2. *While the paper mentions comparisons with other methods, it lacks detailed benchmarking data, such as false positive rates.*

   Response: We appreciate the reviewer's insightful comment regarding the need for more detailed benchmarking data, particularly including false positive rates. In response, we have added a False Positive Rate to the results in Table 2, addressing this specific concern.

   Additionally, to provide a more comprehensive view of the model's performance on imbalanced datasets while addressing other reviewers' comments, we have included AUC and mean of range-based AUC in Table 2. These metrics offer further insights into the detection capabilities of our proposed method.

3. *Comparing diagnosis accuracy for root causes against other interpretable methods could better highlight capabilities.*

   Response: We appreciate the reviewer's suggestion regarding the comparison of diagnosis accuracy for root causes against other interpretable methods. While adapting other interpretable methods for diagnostic purposes, from cited publications, proved challenging due to undisclosed or non-adaptable code, we recognized the importance of addressing this aspect.

To address this concern, we decided to compare our proposed method with DBStream [HB16], an online version of DBScan designed for evolving data streams. It is an unsupervised method with the ability to disregard anomalies and capture clusters of various sizes and densities. It has been successfully used throughout the literature for diagnostic purposes. In [LZLW19], DBScan was used for thermal runaway diagnosis of battery systems in electric vehicles, relying on engineered features reflecting battery performance. Other applications include power transformer fault diagnosis [LSW+20] and fault diagnosis of rolling bearing [LWHL20]. Its usage in these applications, with expert knowledge of the system, involves mapping clusters to root causes, making it a promising method for our comparison.

The comparison was performed on the Controlled Anomalies Time Series (CATS) Dataset, a simulated complex dynamical system with 200 injected anomalies. This dataset is publicly available and suitable for root cause analysis, as disclosed in the dataset's description. We performed two types of comparisons.

Firstly, we assessed clustering performance, a key feature of DBStream. This analysis demonstrated the capability of both methods to separate various groups of anomalies from normal data. Both methods were optimized to maximize the adjusted mutual information score, a metric that showed the best convergence for both methods on dummy data. The results of this comparison are presented in Table 2.

Secondly, we compared the ability to detect root causes. Since DBStream does not inherently provide information about root causes, we introduced expert knowledge a posteriori by mapping clusters to ground truth root causes based on the highest overlap. While this approach is artificial, it facilitated a comparison of the methods in root cause detection, with a slight benefit given to DBStream.

It's important to note that the cluster-to-root cause mapping was also performed during the optimization of DBStream's hyperparameters, resulting in a slightly unfair comparison for our proposed method.

The detailed analysis revealed that our proposed method exhibits a higher precision in detecting root causes compared to DBStream. We have disclosed the results of this comparison in Table 3.

References

[AHMS13] Oren Anava, Elad Hazan, Shie Mannor, and Ohad Shamir. Online learning for time series prediction. *CoRR*, abs/1302.6927, 2013.

[BS17] Dariusz Brzezinski and Jerzy Stefanowski. Prequential auc: properties of the area under the roc curve for data streams with concept

Table 2: Evaluation of AID and DBStream models optimized for adjusted mutual information score on CATS dataset. The best-performing model is highlighted in bold. Perfect clustering achieves 100% in each metric.

Algorithm	AID	DBStream
Adjusted Mutual Info [%]		
Adjusted Rand [%]		
Completeness [%]		
Fowlkes-Mallows [%]		
VBeta [%]		

Table 3: Evaluation of AID and DBStream models optimized for Weighted Precision score on SKAB dataset. The best-performing model is highlighted in bold. Perfect clustering achieves 100% in each metric. Values in brackets represent macro values of the metric.

Algorithm	AID	DBStream
Weighted Precision [%]		
Weighted Recall [%]		
Weighted F1 [%]		

drift. *Knowledge and Information Systems*, 52(2):531–562, Aug 2017.

[DSXS21] Shohreh Deldari, Daniel V. Smith, Hao Xue, and Flora D. Salim. Time series change point detection with self-supervised contrastive predictive coding. In *Proceedings of the Web Conference 2021*, WWW '21, pages 3124–3135, New York, NY, USA, 2021. Association for Computing Machinery.

[HB16] Michael Hahsler and Matthew Bolaos. Clustering data streams based on shared density between micro-clusters. *IEEE Trans. on Knowl. and Data Eng.*, 28(6):1449–1461, jun 2016.

[LSW+20] Yongxin Liu, Bin Song, Linong Wang, Jiachen Gao, and Rihong Xu. Power transformer fault diagnosis based on dissolved gas analysis by correlation coefficient-dbscan. *Applied Sciences*, 10(13), 2020.

[LWHL20] Hai Li, Wei Wang, Pu Huang, and Qingzhao Li. Fault diagnosis of rolling bearing using symmetrized dot pattern and density-based clustering. *Measurement*, 152:107293, 2020.

[LZLW19] Da Li, Zhaosheng Zhang, Peng Liu, and Zhenpo Wang. Dbscan-based thermal runaway diagnosis of battery systems for electric vehicles. *Energies*, 12:2977, 08 2019.

[MBMO16] Igor Melnyk, Arindam Banerjee, Bryan Matthews, and Nikunj Oza. Semi-markov switching vector autoregressive model-based anomaly detection in aviation systems. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1065–1074, New York, NY, USA, 2016. Association for Computing Machinery.

[ZZQ23] Ruiyao Zhang, Ping Zhou, and Junfei Qiao. Anomaly detection of nonstationary long-memory processes based on fractional cointegration vector autoregression. *IEEE Transactions on Reliability*, pages 1–12, 2023.