POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Anomaly Detection in Multivariate Time Series: Comparison of Selected Inference Models and Threshold Definition Methods

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Gioele Verze

Advisor: Prof. Piero Fraternali

Co-advisor: Nicolò Oreste Pinciroli Vago

Academic year: 2021-2022

1. Introduction

Anomaly detection is a technique used to identify observations or events that deviate significantly from normal behavior [3]. Anomaly detection applications range from industrial maintenance to finance and medicine. The feature that unites all these fields is detecting anomalies as soon as they happen to mitigate their impact on the system. Anomaly detection comprises supervised and unsupervised methods. When the training data are explicitly labeled as normal and anomalous, the method is supervised. On the other hand, unsupervised methods learn starting from unlabeled data. Unsupervised anomaly detection methods can detect anomalies in complex systems and can be adapted to identify anomalies in real-time. Moreover, unsupervised anomaly detection can detect new types of anomalies which do not need prior knowledge. Also, the data on which anomaly detection is performed could be of different types, such as images or time series. This work focuses on time series which can be either univariate or multivariate. In the latter case, multiple variables are collected simultaneously over time, providing a better overview of the system than the univariate ones. This work analyzes the impact of different machine learning models and threshold selection in multivariate time series. In particular, it analyzes the behavior of the state-of-the-art methods on SKAB [8] and EXATHLON [7], two datasets commonly used in anomaly detection literature. The results demonstrate that, despite deep-learning methods having quite different performances, the choice of the threshold is fundamental. For this reason, the threshold should be carefully considered when evaluating and comparing anomaly detection methods.

2. Related Work

Anomaly detection is addressed by a wide variety of techniques, including classification-based, nearest neighbors distance-based, clustering-based, statistical, and ensemble approaches.

Distance-based nearest-neighbor techniques compare suitably defined points to their nearest neighbors and flag the ones significantly different from their neighbors as anomalies. Time series data are typically segmented into fixed-length windows encoded in a multidimensional space. The underlying idea is that normal points are close to each other, whereas anomalies are more distant from their nearest neighbors. Examples include KNN, which compares each point to its k nearest neighbors, and LOF, which considers the data density in the area surrounding a point and signals points in low-density areas as anomalous.

Clustering-based approaches project the data onto a multidimensional space and compute the density of resulting clusters. Observations within low-density clusters are considered anomalous.

Statistical and probabilistic methods model the data using a training set and estimate the probability that a test sample belongs to the distribution. Parametric approaches assume that normal data are generated by a known parametric distribution. The anomaly score of a test sample is obtained by performing a statistical hypothesis test. Non-parametric approaches do not assume a density distribution a priori but estimate it from the data with such techniques as histogram-based (HBOS) and kernel function-based models.

Ensemble methods combine multiple approaches using a consensus system. This technique is very effective in datasets that contain different types of anomalies because a certain approach can perform well in finding specific types of anomalies.

Classification-based approaches use a classifier to distinguish between normal and anomalous instances. They can address both one-class and multi-class anomaly detection tasks. Representative examples are neural networks, Bayesian networks, support vector machines, and rule-based systems. Unsupervised classification-based approaches require an anomaly-free training set to create a model of normal behavior. Then, the model is used on the test set, producing an anomaly score for each data point.

The anomaly detection methods, which do not return directly if a sample is anomalous or not, return an anomaly score that, compared with a threshold, shows predicted anomalies. Precisely, the threshold function is to separate anomalous and normal values. There are several approaches to computing a suitable threshold. One common technique consists of calculating the threshold on a validation set and then using it to evaluate each test sample. This is the case of [7], which uses IQR (Inter-Quartile Range), MAD (Median Absolute Deviation), and STD (Standard Deviation), as they are among the most used thresholding techniques. Others set the threshold to the maximum value of the validation anomaly score. Instead of using statistical properties, such as mean or median, others proposed a nonparametric dynamic thresholding method that does not assume a specific underlying distribution of the anomaly scores and is based on a single parameter (z), which is set experimentally.

Other approaches, which are not applicable to online anomaly detection, consist in setting the threshold directly on the test set. Some works compute the threshold as the cut-off value leading to the optimal separation between normal and anomalous data in the test set; others select the threshold as the value that maximizes the F1 score on the test set.

3. Model Design

3.1. Model

The machine learning models analyzed and used can be classified as prediction-based, reconstruction-based, and clustering-based. Prediction-based models focus on, starting from an input sequence, predicting the following sequence according to what the model has learned during the training. In particular, the predicting-based model used in the thesis is:

- LSTM [9] composed of two stacked Long short-term memory layers

Reconstruction-based models instead aim to learn a lower-dimension representation for higher-dimensional data by training the network to capture the most crucial features of the input data. Autoencoders belong to this category. They are a particular neural network composed of three distinct parts: encoder, decoder, and latent space. The encoder is the module that compresses the input data into a representation smaller than the original, while the decoder has the opposite function. Starting from the lower dimension, reconstruct the data to the original size. In the middle, between the encoder and decoder, there is the latent space which contains the compressed knowledge input representation. The autoencoders used are:

- DENSE-AE [5] has the encoder and the decoder composed of fully dense layers;
- CONV-AE [6] has the encoder and the decoder composed of convolutional layers;

- LSTM-AE [4] has the encoder and the decoder composed of LSTM layers;
- VAE [1] has the encoder and the decoder composed of LSTM layers and earns a representation of the input data by modeling latent variables distribution;
- USAD [2] based on two autoencoders trained adversarially.

Further specifications need to be done about VAE. Unlike other autoencoders, its latent space is continuous, regularized, and complete. So, anomaly detection with VAE can be performed by analyzing the latent space distribution using KNN, Isolation Forest, or ReEnc [11].

The last method analyzed belongs to clustering-based anomaly detection and is called ELM-MI [10]. It is based on extreme learning machines and mutual information combined with a dynamic kernel selection method.

3.2. Threshold

This work analyzes four different thresholding techniques applied to an anomaly score obtained by processing a validation set. In the following algorithms, τ indicates the threshold value, s is the input anomaly score, and th_{factor} is a constant.

- Maximum Value (MV) computes the threshold as the maximum anomaly score on the validation set:

$$\tau = max(s)$$

- Standard Deviation (STD) relies on the anomaly score mean and standard deviation. In this case, th_{factor} represents the minimum number of standard deviations to consider a score anomalous. The threshold is:

$$\tau = mean(s) + th_{factor} \cdot std(s)$$

- Median Absolute Deviation (MAD) computes the median of the absolute differences between each anomaly score and the median anomaly score, making this approach less sensitive to outliers than the standard deviation. It is calculated as:

$$\tau = md(s) + 1.4826 \cdot th_{factor} \cdot md(|s - md(s)|)$$

Where $md(s)$ corresponds to the median of score s

- Inter-Quartile Range (IQR) is based on the difference between the 75th percentile (Q_3) and the 25th percentile (Q_1) of the anomaly scores. It is calculated as follows:

$$\tau = Q_3 + th_{factor} \cdot (Q_3 - Q_1)$$

The choice of the best thresholding method depends on the anomaly scores distribution on non-anomalous validation sets. For example, the Maximum Value approach is particularly sensitive to outliers in anomaly scores. For this reason, it is more effective if the anomaly scores distribution associated with normal values does not present outliers. Other methods, such as IQR and STD, are less sensitive to the presence of outliers, as they consider a range depending on the number of standard deviations from the mean or the 75th percentile. The work in [7] also highlights the challenges in selecting an optimal threshold on non-anomalous data, considering the IQR, SD, and MAD approaches.

4. Dataset

The work experiments are performed on two different datasets.

4.1. SKAB

SKAB dataset [8] is a multivariate time series collected from eight different sensors installed on a pump that undergoes several tests on a test bench. The tests simulate different work conditions to record normal behavior and stress the pump to generate anomalies. Each record has a sample rate of one second. It is composed of 34 files that contain anomalies caused by different factors. Each file is divided into three parts:

- Training set, which contains only normal data and is composed of the first 400 values and used to train the neural network model;
- Validation set, which contains only normal data and is used to compute the threshold
- Test set which contains both normal and anomalous data and processed to find anomalies

4.2. Exathlon

Exathlon dataset [7] is a multivariate time series composed of 2,283 features built from recording, throw Spark Monitoring and Instrumentations Interface, the repeated execution of ten different Spark stream processing applications on a

4-node cluster. Each record has a sample rate of one second. Therefore, the dataset could be split into two parts:

- Undisturbed traces which contain only normal data. This portion of the dataset is, in turn, split into a training set (70%), a validation set (15%), and a threshold set (15%). The first two parts are used to train the neural network models, while the latter is used to compute the threshold.
- Disturbed Traces which contain both normal and anomalous data. They composed the test set, the portion of the data processed by models to detect anomalies.

Since the huge amount of data, the number of features is reduced by PCA to 19, and the number of timestamps is reduced by a resampling factor of 15.

5. Results

5.1. SKAB

To evaluate how each model works independently from the threshold value, the AUROC value is calculated for each anomaly score. It measures how well the model separates the positive and negative classes.

METHOD	CONV-AE	VAE	ReEnc	USAD
AUROC	0.94221	0.92289	0.91611	0.90503

METHOD	ELM-MI	LSTM	LSTM-AE	DENSE-AE
AUROC	0.89592	0.87697	0.87045	0.85651

Table 1: AUROC value according to the different neural network models

As Table 1 shows, CONV-AE is the method with the best AUROC value, while DENSE-AE is the worst. This different behavior is related to the score distribution of the two models. Concerning CONV-AE, the score distributions assumes a bimodal distribution where the peaks correspond to normal and anomalous data and only a few score value are in the wrong distribution. On the other hand, the distribution of DENSE-AE is not bimodal, with no visible separation between normal and anomalous data. Moreover, many normal and anomalous scores are mixed, making the separation inaccurate. The same behavior emerges by comparing Figure 1 with AUROC value. Again, the method with the highest AUROC also has the highest F1-Score, while the method with the lowest AUROC also has the lowest F1-Score. Another relevant consideration is the MV thresholding method. For SKAB, it is a technique that works well, being the method with the best F1-Score for VAE+ReEnc, because the MV threshold has a closer value to the STD threshold with th_{factor} equal to 2. This happens since, in the validation score distribution, no values deviate consistently from the mean (i.e., maximum three std).

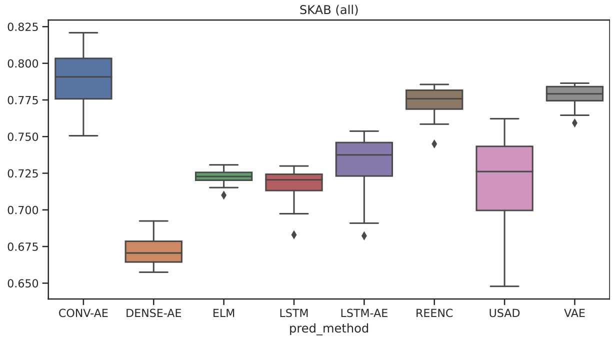

Figure 1: Box plot of F1-Score behavior according to the threshold technique and models

5.2. Exathlon

Also, for the Exathlon dataset, the anomaly score is evaluated independently from the threshold.

METHOD	ELM-MI	CONV-AE	ReEnc	LSTM
AUROC	0.91557	0.89538	0.8880	0.87921

METHOD	USAD	VAE	LSTM-AE	DENSE-AE
AUROC	0.89079	0.85385	0.84996	0.75564

Table 2: AUROC value according to the different neural network models

Table 2 shows that there are methods, like ELM-MI, able to separate anomalies and normal samples well, and methods, like DENSE-AE, that do not work well. Concerning DENSE-AE, the behavior is the same as for SKAB, while ELM-MI behaves differently. Despite having the highest AUROC, the anomaly score has no bimodal distribution. This happens because by design as [10], the score is limited to 0.5, and the anomalies have a score closer to that value.

As Figure 2 and Figure 3 show, comparing them to the AUROC value emerges that, also in this case, the best F1-Score corresponds to the best AUROC. In that case, two different boxplots are shown because the F1-Score changes significantly according to the threshold method. The

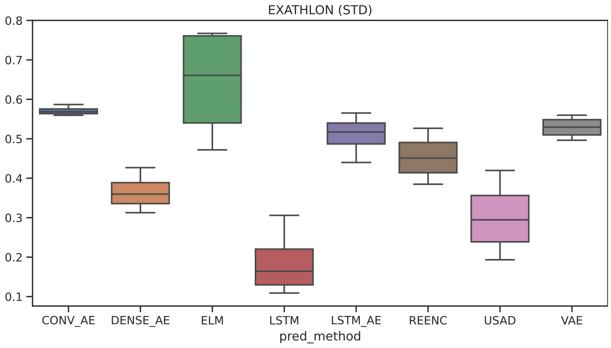

Figure 2: Box plot of F1-Score behavior according STD threshold technique and models

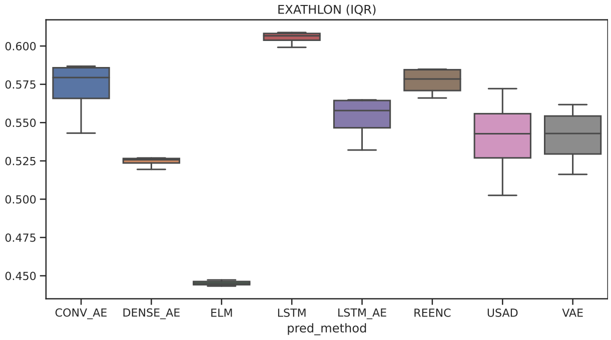

Figure 3: Box plot of F1-Score behavior according IQR threshold technique and models

methods more sensitive to the changing of the threshold technique are ELM-MI and LSTM. Concerning ELM-MI, IQR cannot separate normal and anomalous data well because it generates a too-low threshold. Also, the th_{factor} has no relevant impact, and the ΔF1-Score (difference between maximum and minimum F1-Score) is equal to 0.004. Instead, STD generates a greater threshold, reducing the false positives and obtaining an F1-Score equal to 0.7672. Also, the th_{factor} is very important, having a ΔF1-Score equal to 0.2953. Concerning the LSTM method, the behavior is the opposite. It has good results with IQR and meager results using STD. By comparing the threshold IQR is much less than STD, so STD has too many false negatives.

Concerning the MV technique, what emerges is that it cannot be applied to this dataset. This is because the validation set contains values that deviate consistently from the mean (i.e., 20 or more std). The result is a threshold too high to detect sufficient anomalies.

6. Conclusion

From the experiment results, relevant considerations emerge concerning the different impacts that both models and threshold techniques have on anomaly prediction and, consequently, on evaluation metrics like F1-Score, Precision, and Recall. The effect thresholding techniques depend highly on the anomaly score distribution on which they are computed. If they are calculated on a small validation set and the validation score assumes a nearly uniform distribution, the threshold computed by different techniques is quite similar. Also, the contribution of different th_{factor} is minimal. The result is that the neural network model is the principal choice for better anomaly detection. On the other hand, for huge validation sets, the behavior is the opposite. Being calculated on more value and score distribution containing samples that deviate significantly from the mean, the thresholds assume different values according to the technique used. Moreover, only statistically-based techniques can be used since MV generates inconsistent results. The choice of the threshold factor instead is not so important. It takes values only if the score distribution is limited by a maximum value. The result is that in this type of validation score distribution, the most significant contribution to the F1-Score, Precision, and Recall values corresponds to the choice of the threshold method. This does not mean that the impact of the model is null but lower.

References

[1] Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. *Special lecture on IE*, 2(1):1–18, 2015.

[2] Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A. Zuluaga. Usad: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3395–3404, New York, NY, USA, 2020. Association for Computing Machinery.

[3] Varun Chandola, Arindam Banerjee, and

Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3), jul 2009.

[4] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[5] Gabriel Coelho, Luís Miguel Matos, Pedro José Pereira, André Ferreira, André Pilastri, and Paulo Cortez. Deep autoencoders for acoustic anomaly detection: experiments with working machine and invehicle audio. *Neural Computing and Applications*, 34(22):19485–19499, 2022.

[6] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742, 2016.

[7] Vincent Jacob, Fei Song, Arnaud Stiegler, Bijan Rad, Yanlei Diao, and Nesime Tatbul. Exathlon: A benchmark for explainable anomaly detection over time series. *arXiv preprint arXiv:2010.05073*, 2020.

[8] Iurii D. Katser and Vyacheslav O. Kozitsin. Skoltech anomaly benchmark (skab). `https://www.kaggle.com/dsv/1693952`, 2020.

[9] Benjamin Lindemann, Timo Müller, Hannes Vietz, Nasser Jazdi, and Michael Weyrich. A survey on long short-term memory networks for time series prediction. *Procedia CIRP*, 99:650–655, 2021. 14th CIRP Conference on Intelligent Computation in Manufacturing Engineering, 15-17 July 2020.

[10] Xinggan Peng, Hanhui Li, Feng Yuan, Sirajudeen Gulam Razul, Zhebin Chen, and Zhiping Lin. An extreme learning machine for unsupervised online anomaly detection in multivariate time series. *Neurocomputing*, 501:596–608, 2022.

[11] Chunkai Zhang, Shaocong Li, Hongye Zhang, and Yingyang Chen. Velc: A new variational autoencoder based model for time series anomaly detection. *arXiv preprint arXiv:1907.01702*, 2019.