Adaptive Threshold for Anomaly Detection in ATM Radar Data Streams

Achraf Krim Rahaoui, Théobald de Riberolles, Jiefu Song

Adaptive Threshold for Anomaly Detection in ATM Radar Data Streams

Achraf Krim Rahaoui[1(✉)], Théobald de Riberolles[1], and Jiefu Song[1,2]

[1] Activus Group, 2 Chemin du Pigeonnier, 31100 Toulouse, France
{achraf.krimrahaoui,theobald.deriberolles,jiefu.song}@activus-group.fr
[2] IRIT - Université Toulouse I Capitole, 2 Rue du Doyen Gabriel Marty, 31042 Toulouse Cedex 09, France

Abstract. Intrusion Detection Systems (IDS) are capital instruments for protecting ATM networks against intrusion, and subsequently ensuring the integrity of air traffic. An anomaly detection approach in such systems enables the detection of multiple types of attacks with the aid of a threshold as a criterion for differentiating between normal activity and unusual events in the network. IDS with fixed threshold fail to detect the presence of patterns in the data, thus hampering proper detection ability, and requiring regular human intervention. Detection ability of IDS can be improved by establishing an automated system that recognises pattern shifts in evolving data streams and adjusts the threshold accordingly. Our work focuses on designing an algorithm to recognize the occurrence of new patterns and adjust the threshold consequently for enhanced anomaly detection, whilst offering flexibility for different frameworks and scalability to cope with large data streams. In this article, we present an adaptive threshold approach based on extreme value theory, which aims to automatically detect concept drifts in radar data streams. We evaluate our method in a practical scenario of anomaly detection on time series data collected by air traffic radars across France and show that we can achieve a threefold performance improvement over a standard approach using a fixed threshold.

Keywords: Anomaly detection · Concept drift · Time series

1 Introduction

Surveillance radars for air traffic monitor the behaviour of an aircraft during the course of a flight. Such communications between surveillance radars and aircraft are recorded and transferred over a private network linking the various air traffic management (ATM) entities. Nevertheless, the effort to connect multiple ATM systems, which were previously operating in a closed environment, is resulting in the disruption of previous security features, thereby exposing the entire system to attacks. In order to ensure security in the ATM network, intrusion detection systems (IDS) that can provide a hybrid of misuse and anomaly detection are deployed. Since attacks are not yet widespread in ATM networks, malicious

actions are barely defined, thus an IDS based on misuse detection is unable to detect unknown attacks. As air traffic increases, it is necessary to ensure the reliability and relevance of the detections of all types of attacks. Existing approaches based on a fixed threshold trigger numerous false alarms, resulting in the failure to meet the criterion of relevance of an IDS. Hence, the importance of introducing an alternative approach that breaks away from the classical approach in order to respond to the qualities of an IDS.

We aim to provide a solution centred on concept drift recognition, with improved anomaly detection over a fixed threshold due to the dynamic adjustment of the threshold following the occurrence of drifts. The proposed solution intended to be flexible and scalable, with the potential to be used in different frameworks and to process high volumes of data. In this article, we present an adaptive threshold approach based on extreme value theory (EVT), which aims to automatically detect concept drifts in radar data streams. We evaluate our method in a practical scenario of anomaly detection on time series data collected by air traffic radars across France and show that we can achieve a threefold performance improvement over a standard approach using a fixed threshold. We also present a protocol for processing radar data to perform anomaly detection, as it constitutes a preliminary step in order to asses our method.

We structure this article as follows. Section 2 presents the state of the art and related work. In Sect. 3 we present the VPOT approach and provide a breakdown of our methodology. In Sect. 4 we provide the experimental framework and discuss the outcomes. Finally, in Sect. 5 we review our progress and consider the directions for future work.

2 Related Work

The existing approaches for anomaly detection through machine learning include outlier detection (Lazarevic et al. [1]), classification (Bhuyan et al. [2]) and semi-supervised learning techniques (de Riberolles et al. [3]). In all of the approaches discussed, anomalies are detected via a fixed threshold. The capacity of a fixed threshold is compromised by the presence of fluctuating data, which will often require human intervention to correct the threshold. Considering that under this approach the threshold is computed on the observed data set, calculating the threshold on a relatively small data set will result in a poor extrapolation of the threshold in data streams. It will thus be required to calculate it on fairly large data sets in order to capture the overall behaviour. However, the processing time of this threshold will be significantly long, especially when the data set processed is large, thereby affecting its scalability. An effective approach to ensure that manual intervention is not required to make adjustments, and to reduce computing time through automation, is to employ an adaptive threshold, which is a mechanism that has the ability to recognise the presence of new behaviour patterns in the observed data in order to calculate a threshold on that basis. In order to develop an adaptive threshold, numerous strategies have been presented in the scientific literature. Machine learning related methods are

addressed for detecting concept drift and computing the threshold. Esposito et al. [4] present a method that relies on Cohen's Kappa coefficient as a metric to ascertain the threshold. This method is suitable for classification algorithms as it does not constrain the training of the model being used. However, the drawing of random samples in the process of calculating the threshold risks breaking the time series continuity, resulting in a poor representation of the behaviour of the data. Probabilistic approaches rely on the results of probability theory, Ali et al.[5] study the automation of the threshold used for anomaly detection systems in an effort to improve the detection capability of zero-day attacks in the traffic of a computer network. To address this aim, a recognition algorithm based on Markov chains is used to predict abnormal scores, leading to an a priori threshold that fits these predictions. An extension of the work of Ferragut et al. [6], who reformulate the notion of anomaly based on a probabilistic approach, is proposed by Bridges et al. [7]. In their studies, the distribution of incoming data is assumed to be known and a guideline is suggested to build a threshold –either fixed or automated– in a generic form that would depend on the data flow and the number of alarms allowed.

An interesting insight is brought by combining the using of sliding windows and probabilistic tools. T. Wang et al. [8] present a martingale-based method to learn the regularity of the observed data in a sliding window of variable size and to identify the shifts in the data stream. The threshold is computed according to a global factor that ascertains the confidence level of the detection. On the other hand, H. Wang [9] defines a threshold for incoming observations by conducting a wavelet analysis and the resulting confidence interval obtained by using the Central Limit Theorem over a sequence of data on a sliding window of fixed size. On their part, Clark et al. [10] present a method for detecting concept drifts through a sliding window based approach that relies on a statistical test, with the threshold being adjusted accordingly. Finally, Siffer et al. [11] suggest the use of a threshold derived from the results of the extreme value theory. The threshold is defined from the quantiles of a generalized Pareto distribution, in dependence on the sensitivity chosen to differentiate between normal and abnormal data.

We have discussed several approaches to establish a threshold that can adapt to the nature of the data. In their majority, these approaches rely on statistical tests or sliding windows, assuming that the theoretical distribution of the data is known. Nevertheless, in a practical framework, the nature of data is constantly changing. Hence, making an a priori assumption on the nature of the data or the abnormal scores constrains the computation of the threshold to a single case of the model. After a thorough review of the methods discussed, we will consider the potential of linking a probabilistic approach with a sliding window system to provide a solution with a level of flexibility that facilitates the inclusion of an adaptive threshold in the operational environment of an IDS. Existing approaches focus on either detecting concept drift or on computing the threshold, and are often restricted to specific frameworks. We propose a generic approach involving concept drift detection and robust threshold calculation. With this

insight, we design an automated method that employs sliding windows of variable size, along the introduction of a non-parametric test to detect concept drifts in fluctuating data. To achieve better efficiency, the threshold is obtained by using an extreme value theory approach. On completion, we perform a benchmarking analysis of the approaches VATU [10] and DSPOT [11] with our approach when applied to surveillance radar data from ATM networks.

3 Method

In this section we present our approach VPOT that combines the VATU [10] and DSPOT [11] concepts for developing an adaptive threshold capable of identifying concept drifts. We also establish a guideline for obtaining an anomaly score from radar data, on which our method will be evaluated.

3.1 VPOT Approach

VATU approach addresses the detection of concept drift zones in Gaussian distributions via the z-test. To perform the test, two sliding windows are dedicated for the comparison of the last monitored and new incoming scores. However, the z-test is bounded by its inability to perform on non-Gaussian distributions. To ensure compatibility with scores from different distributions, we have introduced the Kolmogorov-Smirnov test in VPOT, which is applied on the scores in the two sliding windows. In contrast to VATU, where the threshold is calculated by a linear combination of the mean and standard deviation of the scores, a more advanced threshold calculation is possible with DSPOT. This approach based on the EVT, and more accurately on the Peaks Over Threshold (POT) method, allows to compute a threshold without prior knowledge of the distribution of the scores. A crucial aspect is that EVT results, and the subsequent application of the POT method, require the scores to be independent and identically distributed (iid), which is not met in a realistic scenario such as presented in this study. On the case of the scores we generate, they are dependent on each other if the distance between them is less than or equal to the size of the window used for computing the scores. Nevertheless, this dependence weakens as two scores become more distant from each other. By analysing the autocorrelation of the scores (Fig. 1), we observe a decreasing degree of correlation as the lag increases. This outcome, which is consistent with the scoring method, suggests that the dependency between two scores is short-term. Under the condition of short-term dependency, the same results of the EVT can be applied as for the iid variables [12,13].

Moreover, in broad strokes, the POT method is based on the Pickands-Balkema-De Haan theorem [14,15] and is applied for ascertaining the probability of an anomalous event. The method, however, is considered for situations where the scores do not vary considerably, and to ensure this condition, a change of variables $X' = X - M_d$, where M_d is the moving average over the last d scores, is introduced. Finally, the maximum likelihood method is used to ascertain the

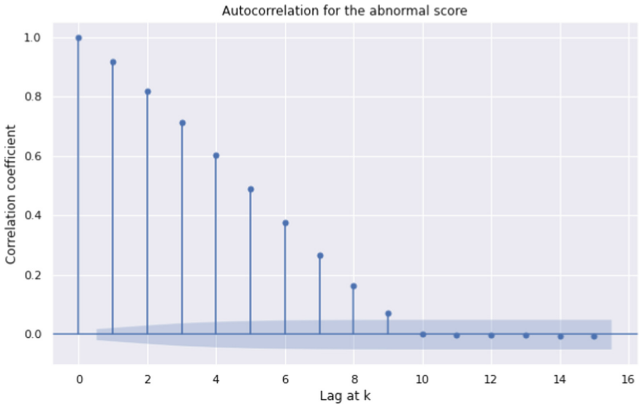

Fig. 1. Autocorrelation of the scores in relation to the lag k

extreme quantile to be used as a threshold for identifying abnormal activity. In DSPOT, the extreme quantile is computed with every new observation. In contrast, in VPOT, every time a drift is detected, the extreme quantile is computed over the scores in a third window that includes both the observed and the incoming scores, thus improving the computational efficiency.

Following a similar procedure as in VATU, we regularly update the scores stored in the third window. The purpose of this update is to prevent a stationary threshold in the case where concept drifts are not present or not properly detected. In the following diagram (Fig. 2), we present the steps of the VPOT algorithm.

3.2 Methodology

In order to evaluate the performance of the discussed methods, we produce scores that capture the degree of abnormality in a sequence of data. In the following paragraphs we will briefly describe the process leading to the setting of our adaptive threshold.

Initially, we collect the features of interest from raw data stored in network packet records. We then proceed to the preparation of the data to be transferred to the autoencoder model for its training. For a brief overview, following on [3], we consider an autoencoder consisting of GRU (Gated Recurrent Units) cells. After concluding the training of the model, we artificially introduce anomalies that represent spoofing attacks in the test samples. The last step in the process consists in comparing the reconstruction provided by the autoencoder with the input data by using a metric –abnormal score– derived from the cosine similarity. Once the scores are computed, we then set up the threshold –fixed or adaptive– to identify anomalies in the testing data sets. On these scores, adaptive threshold algorithms employ sliding windows to identify concept drifts that occur, and set the threshold in accordance with the incoming data. Our protocol is described in Fig. 3.

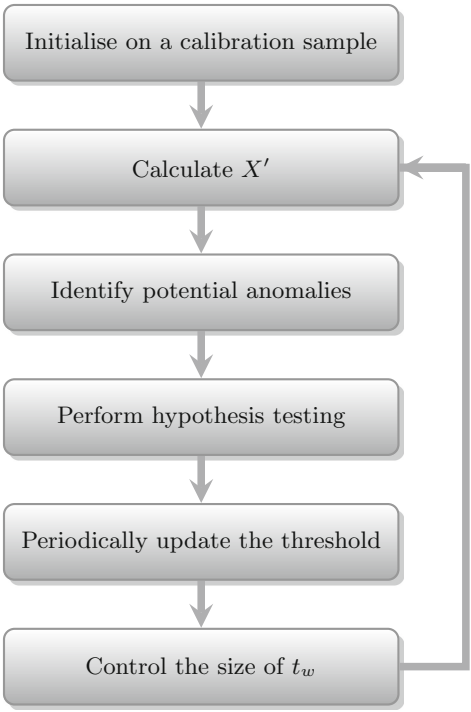

Fig. 2. Algorithm VPOT proceeding

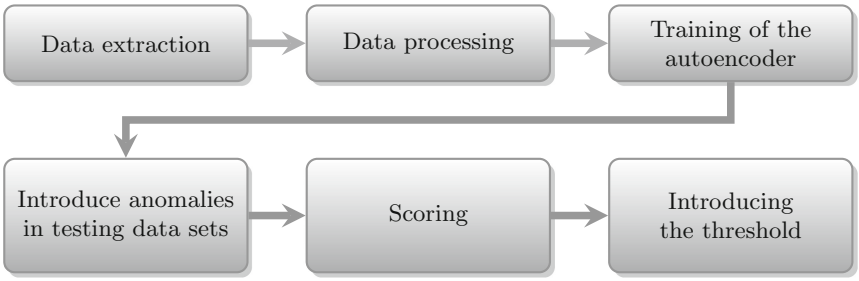

Fig. 3. Stages of anomaly detection using an autoencoder

4 Experimental Assessment

In this section, we will provide our experimental protocol, followed by an overview of the experimental data sets and their corresponding abnormal scores. Finally, we will evaluate and discuss the results of the three methods VATU, DSPOT and VPOT, including their performance with respect to a fixed threshold.

4.1 Experimental Protocol

The purpose of our experiments is twofold. First, to evaluate the ability of our algorithm to identify the areas where concept drift occurs. Our second aim is to assess the effectiveness of the threshold to be adjusted in such a way to detect a maximum of anomalies –criterion of reliability– and to raise the fewest false alarms –criterion of relevance–. For these experiments we formulate the following assumptions. We set ourselves in a realistic framework, in which the data is complex, with the occurrence of different patterns that result in concept drifts in data and subsequently in anomalous scores. The data set used for training the autoencoder consists of real world data from normal air traffic activity. On the other hand, the data sets used for testing include anomalies corresponding to spoofing attacks where information have been altered. In order to measure the scalability of our solution, we chose test samples with size of 3 million records that represent 5% of all records over a 24-h time span.

We perform our experiments in a virtual machine under the following environment: 12 CPUs x Intel Xeon(R) Silver 4216 CPU @ 2.10 GHz, 62.8 GB of RAM, 214.7 GB of disk memory, a 256-bit LLVM 11.0.0 GPU, running on a 64-bit Ubuntu 20.04.2 LTS system. The runtime of the algorithms varies with the sample size. However, we can differentiate the algorithms based on their rapidity. In the following table, we have illustrated the average runtime of the reviewed methods on the larger samples (Table 1).

Table 1. Table of performances

Algorithm	Sample size	Average runtime
VATU	3×10^6	14 s
DSPOT	1×10^6	4 h 36 min 47 s
VPOT	3×10^6	1 min 35 s
Fixed threshold	3×10^6	38 min 54 s

In view of the considerably high execution time for the DSPOT algorithm for large sample sizes, we have opted to rely on the results generated by a one-time execution of the algorithm. Given this constraint, we also choose to benchmark the other algorithms under the same conditions to provide a more fair comparison.

4.2 Data Set

The data at our disposal corresponds to the captured messages sent by surveillance radars for civil aviation air traffic. Raw network capture files (.pcap) contain 4 h of records, with the average size of each file being 700 Mb, which varies depending on the traffic at the time of recording. To facilitate the analysis and processing of such data, we transform the raw capture files into .csv files.

With real world data being used, anomalies are artificially introduced into the data set. Hence, it is not convenient to fix permissible false alarm rates for the framework. Our data sets are built from a collection of information retrieved from different aircraft. This information enables us to identify an aircraft –aircraft address (ACAddr)– and track its position –in polar coordinates (RHO, THETHA)– and route –flight level (FL), calculated ground speed (CGS) and calculated heading (CHdg)–, to identify the radar station transmitting the messages to the network –system identification code (SIC)–, and the timestamp (TS). In view of this background, and the importance of considering the reporting time of the transmitted messages, the data collection under study consists of a multivariate time series. For this paper, we will use a training data set consisting of 12 million records registered over a 24-h period. As data collected since the start of the Covid-19 pandemic constitute an unprecedented scenario, it becomes a serious challenge to characterise a pattern of normal air traffic activity. Our finality being to develop a generic support tool for the air traffic controller, by using data from normal activity, we can assess the benefits of our approach with a real world data set. Therefore, we will be focusing on data collected prior to the pandemic. More precisely, the data set used for training the autoencoder consists of 12 million records, that were retrieved on 24/09/2019. For testing, we use two distinct data sets, each one consisting of 3 million records retrieved on 25/09/2019 and 26/09/2019, over a 4-h time frame corresponding to the peak of air traffic throughout the day in order to gather the more relevant information.

4.3 Benchmarking

Before evaluating the performance of the proposed algorithms, we tuned the parameters of each method to define the optimal setting. The window size is adjusted to allow enough data to compute the threshold. A narrow margin of error is assigned to the test that identifies concept drift. The likelihood of an anomaly occurring is typically low, therefore we opt for large quantiles that enable the recognition of anomalies. We employed the following metrics as benchmarks: precision, recall, accuracy and F1 score. We therefore suggest the following settings for each algorithm (Table 2):

Table 2. Table of optimal settings

Algorithm	Window size	Significance level of the test	Probability (quantile)
VATU	$T = 300$	alpha = 0.05	–
DSPOT	$d = 150$	–	$q = 0.01$ ($z_{0.99}$)
VPOT	$T = 2000$	alpha = 0.05	$q = 0.01$ ($z_{0.99}$)

Taking these settings as a reference, we conducted a benchmark of the performances achieved by the VATU, DSPOT and VPOT algorithms on sub-samples of different sizes, along with the ones achieved by using a fixed threshold. To define the fixed threshold, we selected the threshold s^* that presented the lowest false positive rate (FPR) amongst the thresholds that exhibited a true positive rate (TPR) superior to a certain value δ that is arbitrarily adopted in consideration of the capacity of the autoencoder: $s^* = \underset{s}{\text{argmin}}\{FPR(s); TPR(s) \geq \delta, s \in \mathcal{S}\}$, with $\mathcal{S}$ being the set of the fixed thresholds s used to calculate the ROC curve.

4.4 Observation

After running the algorithms on the scores achieved on subsamples of sizes ranging from 50,000 to 3 million records from the test data sets, we provide the following remarks.

The first remark is that an adaptive threshold provides a better overall performance (Fig. 4 and 6) than the fixed threshold. However, it shows less sensitivity compared to a fixed threshold (Fig. 5 and 7). On the smaller samples, VATU, DSPOT and VPOT have similar performance and sensitivity for the first test data set (Fig. 4 and 5), whereas we perceive a gap in performance and sensitivity between VPOT and DSPOT for the second data set (Fig. 6 and 7). On medium and large samples for both test data sets, both VATU and VPOT maintain the same level of efficiency, with a slightly decreasing sensitivity in spite of a slight decline in the efficiency of VPOT for the second test data set on medium samples. An efficiency decay is noticeable for DSPOT on both data sets. As a result, it appears that overall VPOT shows slightly better results, while VATU and DSPOT exhibit similar performances, all three performing significantly better than a fixed threshold.

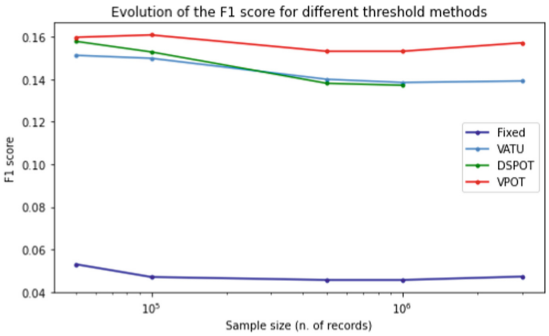

Fig. 4. F1 score for test data set collected on 25/09/2019

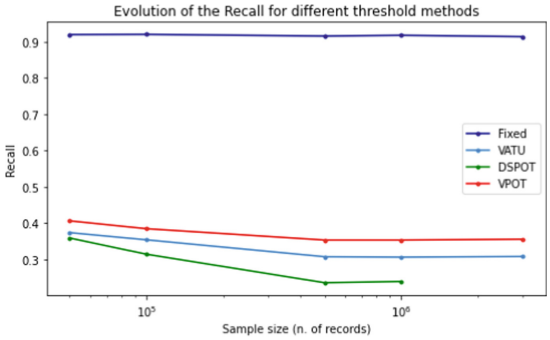

Fig. 5. Recall for test data set collected on 25/09/2019

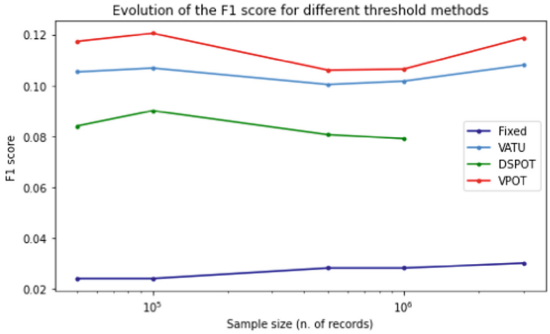

Fig. 6. F1 score for test data set collected on 26/09/2019

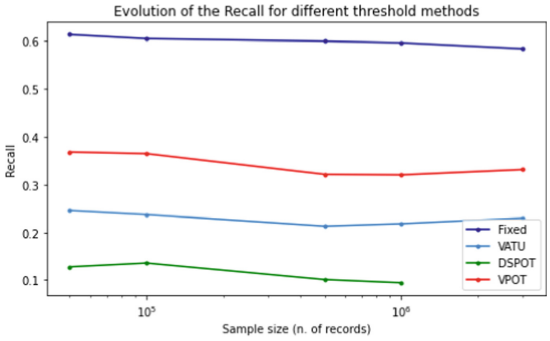

Fig. 7. Recall for test data set collected on 26/09/2019

4.5 Discussion

Firstly, we shall stress that while the fixed threshold ensures a better detection of the introduced anomalies, it triggers frequent false alarms that cause a decrease in precision and a consequent poor performance. Regarding accuracy, the use of

an adaptive threshold produces fewer false alarms, yielding improved precision as compared to the fixed threshold. Both methods VATU and VPOT use the scores within a sliding window (t_w) of reasonably small size to calculate the threshold. This means that the calculated thresholds depend to a lesser extent on the previous scores, ensuring a better fit to the observed patterns, which in turn provides consistency to the methods. In contrast, DSPOT's method of calculating the threshold considers the excesses of all previous observations. As a consequence, a substantial impact is caused by the excesses of data showing a different behavioural pattern and therefore the calculated threshold is increasingly less dependent on the actual data. This allows us to understand the decrease in performance experienced as the sample size increases. We can attribute the enhanced efficiency of the VPOT method to the fact that the EVT provides a threshold that is more adapted to the observed data compared to using a linear combination of the mean and the standard deviation, particularly by using only the excesses within the sliding window. On the second data set, the decay in efficiency experienced by all threshold methods suggests that effectiveness of such methods is higher for data sets corresponding to the following day of the data set used for training the autoencoder.

5 Conclusion and Future Work

In this paper, we have discussed the interest of a developing an adaptive threshold approach for detecting anomalies in ATM networks. To overcome the limitations of an IDS based on a fixed threshold, we have defined an algorithm –VPOT– based on an adaptive threshold approach that provides improved performance, while being consistent, accurate and fast on high volume data streams.

From an overall standpoint, this approach brings us closer to the requirements –reliability and relevance– for an IDS in a wider international ATM network. However to achieve operability in a real life application, certain aspects can be further addressed. In particular, transition areas between data from two separate aircraft are interpreted as anomalies by the autoencoder model, yielding a higher scores that trigger numerous false alarms, thus reducing the performance of the threshold. We therefore contemplate a more extensive handling of concept drift [16] in order to identify transition zones more accurately, and introducing machine learning-based approaches for concept drift detection [17].

In addition, the capacity of the autoencoder is an underlying factor in our protocol, therefore to enhance its efficiency, we intend to conduct continual learning of the model, which will enable us to yield more accurate scores.

References

1. Lazarevic, A., Ertöz, L., Kumar, V., Ozgur, A., Srivastava, J.: A comparative study of anomaly detection schemes in network intrusion detection. In: SDM 2003, May 2003, vol. 3 (2003). https://doi.org/10.1137/1.9781611972733.3
2. Bhuyan, M.H., Bhattacharyya, D.K., Kalita, J.K.: Network anomaly detection: methods, systems and tools. IEEE Commun. Surv. Tut. **16**(1), 303–336 (2014). https://doi.org/10.1109/SURV.2013.052213.00046

3. de Riberolles, T., Song, J., Zou, Y., Silvestre, G., Larrieu, N.: Characterizing radar network traffic: a first step towards spoofing attack detection. In: 2020 IEEE Aerospace Conference, pp. 1–8 (2020). https://doi.org/10.1109/AERO47225.2020.9172292
4. Esposito, C., Landrum, G.A., Schneider, N., Stiefl, N., Riniker, S.: GHOST: adjusting the decision threshold to handle imbalanced data in machine learning. J. Chem. Inf. Model. **61**(6), 2623–2640 (2021). https://doi.org/10.1021/acs.jcim.1c00160
5. Ali, M.Q., Al-Shaer, E., Khan, H., Khayam, S.A.: Automated anomaly detector adaptation using adaptive threshold tuning. ACM Trans. Inf. Syst. Secur. **15**(4), 1–30 (2013). https://doi.org/10.1145/2445566.2445569
6. Ferragut, E., Laska, J., Bridges, R.: A new, principled approach to anomaly detection. In: 2012 11th International Conference on Machine Learning and Applications, December 2012, vol. 2, pp. 210–215 (2012). https://doi.org/10.1109/ICMLA.2012.151
7. Bridges, R., Jamieson, J., Reed, J.: Setting the threshold for high throughput detectors: a mathematical approach for ensembles of dynamic, heterogeneous, probabilistic anomaly detectors. In: 2017 IEEE International Conference on Big Data (Big Data), December 2017, pp. 1071–1078 (2017). https://doi.org/10.1109/BigData.2017.8258031
8. Wang, T., Lu, G.-L., Liu, J., Yan, P.: Adaptive change detection for long-term machinery monitoring using incremental sliding-window. Chin. J. Mech. Eng. **30**(6), 1338–1346 (2017). https://doi.org/10.1007/s10033-017-0191-4
9. Wang, H.: Anomaly detection of network traffic based on prediction and self-adaptive threshold. Int. J. Fut. Gener. Commun. Netw. **8**(6), 205–214 (2015). https://doi.org/10.14257/ijfgcn.2015.8.6.20
10. Clark, J., Liu, Z., Japkowicz, N.: Adaptive threshold for outlier detection on data streams. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), pp. 41–49 (2018). https://doi.org/10.1109/DSAA.2018.00014
11. Siffer, A., Fouque, P.A., Termier, A., Largouët, C.: Anomaly detection in streams with extreme value theory. In: KDD 2017, August 2017, pp. 1067–1075 (2017). https://doi.org/10.1145/3097983.3098144
12. Leadbetter, M., Lindgren, G., Rootzén, H.: Extremes and Related Properties of Random Sequences and Processes. SSS, Springer, New York (1983). https://doi.org/10.1007/978-1-4612-5449-2
13. Poon, S.-H., Rockinger, M., Tawn, J.: Modelling extreme-value dependence in international stock markets. Stat. Sin. **13**(4), 929–953 (2003). https://doi.org/10.2139/ssrn.302961. Institute of Statistical Science, Academia Sinica
14. Balkema, A.A., de Haan, L.: Residual life time at great age. Ann. Probab. **2**(5), 792–804 (1974). https://doi.org/10.1214/aop/1176996548
15. Pickands III, J.: Statistical inference using extreme order statistics. Ann. Stat. **3**(1), 119–131 (1975). https://doi.org/10.1214/aos/1176343003
16. Hoens, T.R., Polikar, R., Chawla, N.V.: Learning from streaming data with concept drift and imbalance: an overview. Prog. Artif. Intell. **1**(1), 89–101 (2012). https://doi.org/10.1007/s13748-011-0008-0
17. Harries, M., Horn, K.: Detecting concept drift in financial time series prediction using symbolic machine learning (July 1996)