ISSN 2219-9365

https://doi.org/10.31891/2219-9365-2024-80-15
UDC 004

PETLIAK Nataliia
Khmelnytskyi National University
https://orcid.org/0000-0001-5971-4428
e-mail: npetlyak@khmnu.edu.ua
BILETSKYI Kostiantyn
Khmelnytskyi National University
e-mail: biletskyik@khmnu.edu.ua
ZASTAVNA Yana
Khmelnytskyi National University
e-mail: zastavna@khmnu.edu.ua

APPROACH TO DETECTION OF ANOMALOUS NETWORK TRAFFIC USING LOF AND HBOS ALGORITHMS

The article is devoted to the problem of detecting anomalies in modern computer networks, which is one of the main threats to cyber security. With the development of Internet technologies, the number of devices and the volume of network traffic are constantly increasing, which leads to an increase in the risk of various cyber threats, such as DDoS attacks, zero-day attacks, and exploitation of protocol vulnerabilities. Abnormal network traffic can result from malicious activity and technical malfunctions, such as configuration errors or hardware failures. Specialised algorithms and methods of analysing large volumes of data are used to detect such threats. The paper considers the main methods of detecting anomalies in network traffic, including classical approaches and modern deep and machine learning methods. Special attention is paid to the efficiency of using methods based on convolutional neural networks, long-term memory and their combinations to detect anomalies. An analysis of the disadvantages and advantages of various approaches to detecting anomalous traffic, such as high computational requirements and the complexity of setting up models, is performed. Still, their effectiveness in analysing large volumes of data is noted. One of the main methods used for anomaly analysis is the local outlier algorithm, which compares the density of objects with their neighbours, allowing for the detection of anomalies in regional segments of the data. Another method is histogram-based outlier estimation, which is faster and more efficient using one-dimensional histograms for each variable. The work also explores the application of unsupervised machine learning methods, which allows for analysing network traffic in real time without the need for prior labelling of data. The article also considers the prospects of further testing the proposed methods in real networks. The combined use of LOF and HBOS balances anomaly detection accuracy and data processing speed, essential to ensure continuous system operation in high-load networks. The implementation of similar solutions in actual conditions requires further research, particularly regarding optimising the use of computing resources and adapting methods to the specific conditions of the network environment. Thus, the paper presents a thorough analysis of modern approaches to detecting anomalies in network traffic and substantiates the feasibility of their application in actual conditions to increase the effectiveness of cyber security.

Keywords: network traffic, anomalies, anomaly detection, local emission factor, estimation of emissions based on histogram

ПЕТЛЯК Наталія. БІЛЕЦЬКИЙ Костянтин, ЗАСТАВНА Яна
Хмельницький національний університет

ПІДХІД ДО ВИЯВЛЕННЯ АНОМАЛЬНОГО МЕРЕЖЕВОГО ТРАФІКУ З ВИКОРИСТАННЯМ АЛГОРИТМІВ LOF ТА HBOS

Стаття присвячена проблемі виявлення аномалій у сучасних комп'ютерних мережах, яка є однією з основних загроз кібербезпеці. З розвитком Інтернет-технологій кількість пристроїв і обсяг мережевого трафіку постійно зростає, що призводить до збільшення ризику різноманітних кіберзагроз, таких як DDoS-атаки, атаки нульового дня, використання вразливостей протоколів. Аномальний мережевий трафік може бути наслідком зловмисної діяльності чи технічних збоїв, наприклад помилок конфігурації або апаратних збоїв. Для виявлення таких загроз використовуються спеціалізовані алгоритми та методи аналізу великих обсягів даних. У статті розглянуто основні методи виявлення аномалій у мережевому трафіку, включаючи класичні підходи та сучасні методи глибокого та машинного навчання. Особливу увагу приділено ефективності використання методів на основі згорткових нейронних мереж, довготривалої пам'яті та їх комбінацій для виявлення аномалій. Проведено аналіз недоліків та переваг різних підходів до виявлення аномального трафіку, таких як високі обчислювальні вимоги та складність налаштування моделей. Проте відзначається їхня ефективність при аналізі великих обсягів даних. Робота досліджує застосування методів неконтрольованого машинного навчання, що дозволяє аналізувати мережевий трафік у реальному часі без необхідності попереднього маркування даних. У статті також розглянуто перспективи подальшої апробації запропонованих методів у реальних мережах. Комбіноване використання LOF і HBOS збалансовує точність виявлення аномалій і швидкість обробки даних, необхідну для забезпечення безперервної роботи системи в мережах з високим навантаженням.

Ключові слова: мережевий трафік, аномалії, виявлення аномалій, локальний коефіцієнт викиду, оцінка викидів на основі гістограми

INTRODUCTION

Abnormal traffic in modern computer networks represents one of the main threats to their security and stable operation. The constant growth of the number of connected devices and the continuous development of

Internet technologies lead to increased traffic that circulates daily in global networks. Along with the increase in traffic, cyber threats also increase [1-2]. From classic DDoS attacks to more sophisticated incidents such as zero-day attacks and exploitation of protocol vulnerabilities, these threats can cause significant damage to both private companies and government institutions. However, abnormal traffic can signal technical malfunctions (for example, configuration errors or hardware failures) and not just malicious actions [3]. In cybersecurity, such deviations often signal attempts to penetrate the network, attacks on services or systems, or spread malicious software. Such actions require rapid identification and neutralisation of threats to minimise losses.

Tools for detecting anomalous traffic are essential to any modern network protection system [4]. They allow you to analyse network traffic behaviour in real-time and detect deviations from standard work patterns. Detecting anomalies, particularly those related to malicious activities, requires sophisticated algorithms to analyse large volumes of data and consider various factors, including the temporal characteristics of traffic, its volume and sources. It should be noted that external attacks and internal threats, such as the compromise of legitimate users or abuse of access rights, can cause abnormal traffic. This makes traffic monitoring and anomaly detection an essential aspect of ensuring cyber security and the uninterrupted operation of network systems and also highlights the need for increased research and development in this area to create more effective and reliable methods of protection against new and more sophisticated cyber threats.

ANALYSIS OF THE LATEST RESEARCH

The paper [5] presents the Data-Oriented Control Intrusion Detection System (DOC-IDS) model for extracting features and detecting anomalies in network traffic using deep learning. The main feature of this model is the integration of the components of a one-dimensional convolutional neural network (1D CNN) and an autoencoder, which allows one to simultaneously extract critical features from traffic data and detect anomalous behaviours. The model can process large volumes of network packet data, providing high accuracy of threat detection thanks to the analysis of complex interrelationships between bytes. Using different types of loss to minimise reconstruction errors and improve classification ability is also a strong advantage of the model. Among the disadvantages of DOC-IDS, one can note its complexity in setting up and the need for enormous computing resources for practical model training. In addition, the model depends on high-quality training data, and its performance may need to improve in cases where insufficiently representative datasets are used.

The study [6] presents a one-class Long Short-Term Memory (OC-LSTM) method for detecting anomalies in large-scale networks. The main advantage of this approach is its ability to train hidden layer features specifically for the anomaly detection task, unlike hybrid methods that use pre-trained models or autoencoders. OC-LSTM uses a loss function similar to OC-SVM, which allows for more flexible solutions for non-linear boundaries between normal and abnormal data. The peculiarity of OC-LSTM is its end-to-end approach to learning without the need to use additional algorithms for feature selection. However, the complexity of optimising the loss function, which is non-convex, complicates the search for optimal solutions.

The article [7] discusses the methodology for unsupervised detection of anomalies in network traffic based on the iterative process of anomaly assessment. The main feature of this approach is the use of two stages of anomaly assessment, which allows an increase in detection accuracy without using labels for model training. The method was tested on the publicly available datasets IDS2018 and DoHBrw, which allowed us to verify its effectiveness under different abnormal traffic conditions. The technique can provide high accuracy even in cases with limited or no training labels. This is achieved through a multi-functional approach to anomaly analysis that considers temporal and statistical traffic characteristics. In addition, using a self-learning mechanism contributes to the gradual improvement of results and allows the detection of more complex anomalies. However, the method's effectiveness decreases with the increase in the share of anomalous traffic since the assumption of the superiority of regular traffic is no longer supported. In addition, for some types of attacks, such as DoS, the method must show more satisfactory results due to their high share in the total traffic, leading to false positive detections.

The research presented in the article [8] is devoted to the application of deep learning in the field of network security, in particular for intrusion detection. The work proposed a model based on convolutional neural networks (CNN-Focal), which uses the Focal Loss function to optimise work with unbalanced data sets. This approach helps to improve the accuracy of attack detection and increase the overall resilience of the model to new types of threats. The main feature of this approach is the use of small convolution kernels, which reduce the number of unnecessary characteristics and increase the performance of the model. In addition, applying a dropout layer prevents the model from being overtrained, and softmax regression is used for multiclass classification. However, significant computational complexity due to the large number of layers and parameters requires powerful hardware resources for training and testing the model.

The study [9] proposes a model for detecting anomalies in network traffic based on a combination of the K-means algorithm and active learning (ALM). The feature of this model is a two-step process, which includes selecting essential features using the Pearson correlation coefficient and the LightGBM algorithm and classifying anomalies based on the K-means method, which allows you to separate normal and abnormal traffic effectively. Despite significant advantages, the model has certain limitations. First, the K-means method depends on the correct

choice of the number of clusters, which can affect the final classification results. Second, the process of diffusion of results through active learning can be computationally complex, which increases resource requirements for processing large volumes of data.

In [10], a technique for detecting anomalies in network traffic based on bilateral long-term memory (BiLSTM) and the mechanism of attention (Attention) is proposed. A feature of this model is its ability to perform two-stage feature extraction from network traffic. First, a feature extraction is performed using BiLSTM, which allows sequential data analysis, taking into account information from both previous and subsequent elements of the sequence. Next, the attention mechanism is used for secondary feature extraction, giving more weight to essential elements and allowing the model to focus better on crucial traffic characteristics. This increases the accuracy of detecting anomalies in the data. One of the main advantages of this approach is the ability to reduce the number of false positives due to the efficient processing of similar traffic features. With the help of the attention mechanism, the model can better focus on the most essential characteristics, which reduces the probability of incorrect classifications. In addition, BiLSTM integration allows the model to work effectively with sequential data, which is necessary for network traffic analysis, where data consistency is often critical. The use of BiLSTM and the attention mechanism significantly increases the computational complexity of the model, which may require additional resources to process large volumes of data in real-time. In addition, the accuracy of the model depends on the quality of data preprocessing, including the process of normalisation and removal of irrelevant features, which is an essential stage of data preparation for the effective functioning of the model.

Research [11] proposes a real-time network traffic anomaly detection methodology based on deep learning, precisely a combination of CNN and LSTM. This combination makes it possible to analyse a large volume of constantly changing data effectively and ensure anomaly detection accuracy in actual network conditions. The main feature of the approach is the ability of models to effectively process network traffic flows, extracting from them essential features and spatio-temporal dependencies. One of the advantages of using CNN-LSTM is the ability of the model to learn on large-scale data, capturing complex spatial and temporal traffic patterns. This provides significantly higher accuracy compared to traditional machine learning methods. In addition, the study emphasises the importance of processing data streams in real-time with minimal delays, which is essential for promptly detecting cyber threats. Model optimisation through transfer learning, model compression, and parallelisation can reduce computational costs and improve performance in resource-constrained environments like mobile or IoT devices. However, the method has certain limitations. First, many traffic anomalies can lead to class imbalance problems, making it difficult to train the model. Second, significant computational requirements can hinder deploying such solutions in systems with limited resources.

AN APPROACH TO DETECTING ANOMALOUS TRAFFIC

The approach to detecting anomalies in network traffic, as shown in Figure 1, is structured in three steps: data extraction, anomaly detection, and alert verification. The three-step structure allows for the consistency of the network traffic analysis process from the initial data collection to the final verification for compliance with security policies. The first data extraction stage uses a packet analyser to collect and filter information about network packets. This provides the necessary basis for further analysis of network activity. The second stage is the anomaly detection process, based on algorithms for processing the collected data. This step allows the detection of anomalies in network traffic using unsupervised techniques such as local outlier factor (LOF) and histogram-based outlier estimation (HBOS).

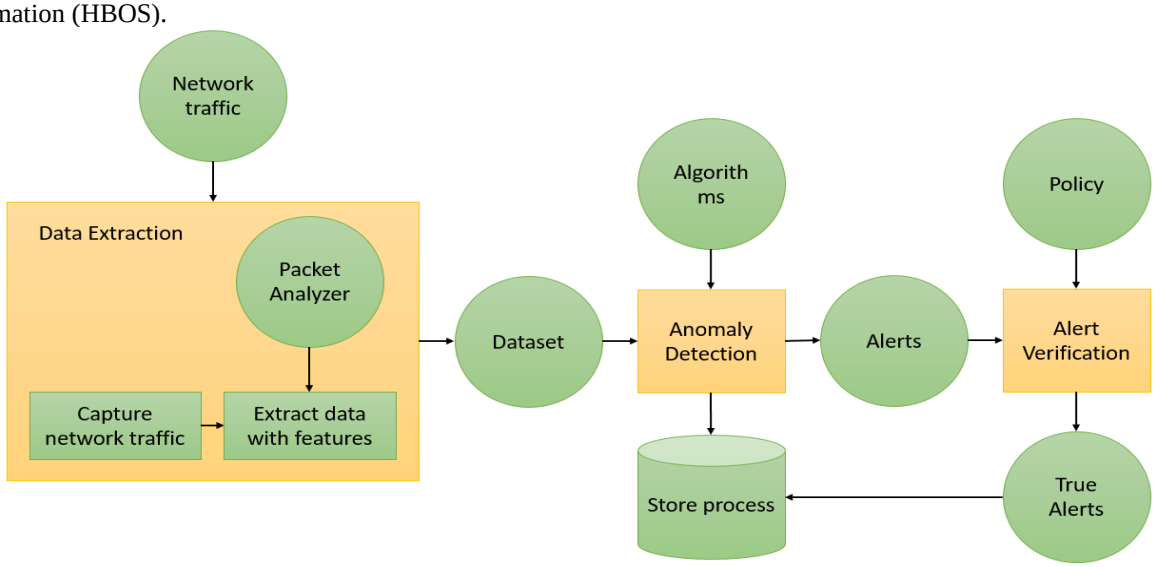

Fig.1. Steps of the proposed approach

The choice of LOF and HBOS algorithms for detecting anomalies in network traffic is justified by their ability to work effectively with different data types and conditions often found in natural network environments. Both algorithms belong to unsupervised machine learning methods, which allow them to be used in cases where there is no prior information about regular or abnormal instances, as is often the case when analysing network traffic. The LOF algorithm is ideal for detecting local anomalies when the density of objects varies in different parts of the data set. This is especially important for analysing network traffic because different parts of the network may have different standard behaviour patterns. LOF compares a point's local density with its neighbours' density, which allows the detection of objects that appear anomalous only relative to their local surroundings, not the entire data set. For example, in cases where individual network segments have specific properties (such as different load levels or traffic types), LOF allows you to recognise anomalies that are not apparent globally. The main advantage of LOF is that it adapts to different local data characteristics, making it flexible and efficient for dynamic environments. The HBOS algorithm, in turn, is distinguished by its ability to work with large volumes of data due to using one-dimensional histograms to estimate the frequency of values of individual features. This makes it fast and productive compared to methods that require multivariate analysis. In the case of network traffic, where the number of parameters can be significant, and changes in individual characteristics, such as IP address or packet size, can indicate an anomaly, HBOS provides an efficient way to estimate the probability of such changes. Using histograms allows the algorithm to automatically adapt to the data distribution and quickly detect deviations even in large sets. Another advantage of HBOS is its ability to process features independently of each other, which simplifies the analysis process in cases where the relationships between features are not critical.

Thus, the combined use of LOF and HBOS makes it possible to balance in-depth analysis of local relationships in the data and fast processing of large arrays of information. LOF provides anomaly detection in complex and heterogeneous environments where data density varies. In contrast, HBOS delivers high speed and scalability, which are critical factors when working with large network data sets. This approach allows not only the accuracy of anomaly detection to be improved but also the optimisation of the use of resources, which is essential in environments with high traffic intensity.

The third stage is notification verification based on access control policies. Using specialised systems to check access rights, the system determines whether detected anomalies correspond to permitted actions according to established security rules. This step ensures the integration of anomaly detection results with existing security policies, which helps avoid false positive alerts and increases the overall effectiveness of the threat detection system.

The process of extracting data is an essential step in detecting anomalies in network traffic. First, collecting and processing network packets allows you to obtain a basic set of data for further analysis. Traffic monitoring is done using the Wireshark tool, enabling you to capture and analyse protocols such as IP, TCP, or UDP. Packet Analyzer provides comprehensive information about protocol type, packet size, IP addresses, and other critical network traffic characteristics essential for identifying potential threats or anomalies. The second stage of data mining is converting the collected information into a format suitable for analysis. This includes the selection of relevant features (features) for model building. These features include source and destination IP address, packet size, delay time, protocol type, and other parameters. It is important to correctly select the features because excessive irrelevant characteristics can complicate further analysis and detection of anomalies. In unsupervised machine learning for anomaly detection, feature selection becomes critical for model accuracy and efficiency.

The transformed data set with all relevant features is the basis for the next step — direct anomaly detection. This step identifies suspicious activities or anomalous behaviour in network traffic that may signal potential threats such as intrusions or malicious attacks. Thus, the quality of the extracted data and the selection of the correct signs of direction affect the effectiveness of the entire anomaly detection system.

After the network traffic capture phase is completed, the main task is to detect anomalies in the data. Unsupervised anomaly detection methods are used in this work since the data do not have labels describing normal or abnormal behaviour. The lack of prior knowledge about the normal state of the system requires the determination of a threshold that separates normal and abnormal behaviour based on statistical indicators or other methods of analysis.

Expected behaviour is defined by a baseline that serves as a benchmark for comparison. Any observations that deviate significantly from this line are considered anomalous or outliers. Each anomaly detection method has its approach to outlier estimation, such as statistical models or data density estimation methods. It is important to note that the right choice of threshold plays a crucial role: a too-high threshold can lead to missing abnormalities, while a too-low threshold can cause many false positives.

After processing all instances of the data set with the anomaly detection method, each instance receives a new attribute. This outlier score indicates the probability that the instance is anomalous. Anomaly detection is often done using machine learning algorithms, such as clustering methods or autoencoders, which efficiently identify anomalous patterns even in large datasets. Thus, the defining aspect of the process is the correct setting of models and thresholds to achieve optimal results.

The system's last stage is verifying notifications, which is based on the analysis of received anomalous cases. Alerts are generated based on received outlier scores that exceed a given threshold. These alerts indicate

suspicious or anomalous activity that needs to be checked through the access control system. We use access control policies and queries to verify that a user can access specific resources. The notification verification process consists of two parts. The first is the creation of queries that are generated based on the alerts generated as a result of the analysis of anomalies. These requests conform to the XACML format and reflect the access policies set in the system. The second part is verification, during which the received requests are checked using access control tools. Verification allows you to confirm whether an alert is a real threat or a false positive. The response of the access control system is based on policies: it can allow or deny access depending on the relevant conditions. This process is crucial to complete the alert verification phase and ensure the information system's security.

CONCLUSIONS

The article emphasises the importance of implementing algorithms for detecting anomalies in network traffic, which allows the detection of both technical malfunctions and potential threats in the form of cyber attacks. The advantages of using machine and deep learning methods in combination with classical methods to increase the effectiveness of network protection are analysed. Further testing of the proposed solutions on real networks will allow us to evaluate their performance in practical conditions, which is the next step in the research. Approbation involves integrating the proposed methods into cyber protection systems for operational real-time traffic monitoring.

References

1. Danial Javaheri, Saeid Gorgin, Jeong-A Lee, Mohammad Masdari, Fuzzy logic-based DDoS attacks and network traffic anomaly detection methods: Classification, overview, and future perspectives, Information Sciences, Vol. 626, 2023, pp. 315-338, doi: 10.1016/j.ins.2023.01.067

2. Xueyuan Duan, Yu Fu, Kun Wang, Network traffic anomaly detection method based on multi-scale residual classifier, Computer Communications, Vol. 198, 2023, pp. 206-216, doi: 10.1016/j.comcom.2022.10.024

3. Haiping Lin, Chengwen Wu, Mohammad Masdari, A comprehensive survey of network traffic anomalies and DDoS attacks detection schemes using fuzzy techniques, Computers and Electrical Engineering, Vol. 104, Part B, 2022, doi: 10.1016/j.compeleceng.2022.108466.

4. Ibrahim Juma, Gajin Slavko, Entropy-based network traffic anomaly classification method resilient to deception, Computer Science and Information Systems, Vol. 19, No. 1, 2022, pp. 87- 116

5. Yoshimura N, Kuzuno H, Shiraishi Y, Morii M, DOC-IDS: A Deep Learning-Based Method for Feature Extraction and Anomaly Detection in Network Traffic, Sensors, Vol. 22, 2022, doi: 10.3390/s22124405

6. Li Y, Xu Y, Cao Y, Hou J, Wang C, Guo W, Li X, Xin Y, Liu Z, Cui L, One-Class LSTM Network for Anomalous Network Traffic Detection, Applied Sciences, Vol. 12, 2022, doi: 10.3390/app12105051

7. Ping G, Zeng T, Ye X, Unsupervised network traffic anomaly detection based on score iterations, Journal of Tsinghua University (Science and Technology), Vol. 62, No. 5, pp. 819-824, doi: 10.16511/j.cnki.qhdxxb.2021.21.045

8. F. Zhao, H. Li, K. Niu, J. Shi, R. Song, Application of Deep Learning-Based Intrusion Detection System (IDS) in Network Anomaly Traffic Detection, Applied and Computational Engineering, Vol. 86, 2024, pp. 250–256, doi: 10.54254/2755-2721/86/20241604.

9. N. Liao, X. Li, Traffic Anomaly Detection Model Using K-Means and Active Learning Method, Int. J. Fuzzy Syst, 2024, pp. 2264–2282, doi: 10.1007/s40815-022-01269-0

10. Pan Chengsheng, Li Zhixiang, Yang Wensheng, Cai Lingyun, Jin Aixin, Anomaly Detection Method of Network Traffic Based on Secondary Feature Extraction and BiLSTM-Attention, Journal of Electronics & Information Technology, Vol. 45, No. 12, 2023

11. Tamilselvan Arjunan, Real-Time Detection of Network Traffic Anomalies in Big Data Environments Using Deep Learning Models, International Journal for Research in Applied Science and Engineering Technology, Vol. 12, No. 9, 2024, doi: 10.22214/ijraset.2024.58946.