UDC 519.6 *http://dx.doi.org/10.30970/vam.2024.33.12791*

INTELLIGENT MONITORING FOR DISTRIBUTED SYSTEMS: LEVERAGING MACHINE LEARNING TO DETECT AND ADAPT TO EVOLVING CYBER THREATS

R. Karpiuk

Ivan Franko National University of Lviv,
1, Universytetska str., 79000, Lviv, Ukraine
e-mail: roman.karpiuk@lnu.edu.ua

Ensuring the stability, performance, and security of distributed computer systems is a critically important task in modern cybersecurity. This paper explores how advanced monitoring systems leveraging machine learning can provide real-time anomaly detection and proactive adaptation. Key challenges include handling large data volumes, architectural complexity, and reducing false positives that burden cybersecurity analysts. The proposed approach integrates centralized data aggregation tools such as ELK and Splunk, machine learning models like Random Forest and DensityFunction for anomaly detection, and scalable microservices architectures deployed on elastic cloud platforms. Additional enhancements include regular model retraining, dynamic threshold adjustments, and automated alerting to improve the detection of evolving cyber threats.

Key words: distributed computer systems, anomaly detection, cybersecurity, machine learning, real-time monitoring, false positives, scalability, microservices architecture, elastic cloud platforms, threat detection, system logs, network traffic analysis, intrusion detection systems (IDS), model retraining, adaptive monitoring systems.

1. Introduction

Ensuring the stability, performance, and security of distributed computer systems is a critically important task in the modern digital world. With the increasing complexity of information infrastructures, data volumes, and heightened risks of cyber threats, there arises a need for monitoring systems that not only detect issues at early stages but also proactively adapt to changes in the environment. In the field of cybersecurity, particular attention is given to anomaly detection, which can indicate potential threats, as well as minimizing false positives and negatives, which directly affect the efficiency of cybersecurity teams. The focus will be on how scalability, availability, flexibility, and the integration of machine learning enable addressing complex challenges in today's digital environment.

2. Model Problem

Distributed computer systems today form the backbone of many businesses, government organizations, and critical infrastructures. Along with technological advancements, new challenges emerge that complicate the task of ensuring their stable operation. One of the key challenges is the growing volume of data. This pertains not only to its storage but also to its analysis for anomaly detection, which may indicate potential threats. For instance, in large networks with thousands of nodes, even a slight delay in threat detection can lead to significant financial losses or data compromise. Additionally, the architectural complexity of distributed systems necessitates the use of adaptive monitoring approaches.

Systems must not only monitor the performance of individual components but also assess their interactions. This is especially crucial in cybersecurity, where attacks can be multi-vector, such as denial-of-service (DDoS) attacks or intrusions exploiting architectural vulnerabilities. Another significant challenge is the need to reduce false positive rates in anomaly detection systems. A high rate of such detections places excessive strain on cyebrsecurity operations center (CSOC) analysts, reducing efficiency and increasing the risk of missing actual threats.

3. Main Research

To build a high-performance monitoring system, it's crucial to focus on the integration of various components that work together to collect, process, and analyze data efficiently. The foundation starts with identifying the right data sources and ensuring that the system can handle large volumes of information in real-time.

When it comes to gathering data, we need to track essential metrics like network traffic, system logs, and security event logs. Network traffic analysis, for instance, helps detect anomalies in packet sizes or unexpected access from unusual IP addresses. System logs, on the other hand, offer a closer look at potential threats by flagging failed login attempts or system errors, which might signal an ongoing attack. Security event logs can also be integrated from intrusion detection systems (IDS) and firewalls, providing a broader view of suspicious activity. All of these data sources should converge in a centralized logging system, such as ELK (Elasticsearch, Logstash, Kibana) or Splunk, enabling real-time aggregation and visualization.

In addition to system logs, integrating security event logs from intrusion detection systems (IDS), firewalls, and antivirus solutions helps provide a more complete view of system security. Each event can be assigned a severity score, which could be represented as:

$$\text{SeverityScore} = \sum_{i=1}^{m} w_i \cdot \text{EventImpact}_i$$

where w_i is the weight assigned to each event, and EventImpact_i reflects the severity of that event.

Once the data is collected, machine learning models are essential for analyzing it effectively, particularly for detecting anomalies and predicting potential threats. For known threats, supervised machine learning algorithms, like Random Forest, can classify traffic as benign or malicious based on historical attack data. The model's effectiveness is often measured by the accuracy of its predictions:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP represents true positives (correctly detected attacks), TN represents true negatives (correctly identified benign traffic), FP are false positives, and FN are false negatives.

For detecting unknown threats, DensityFunction are used in unsupervised learning. These models do not require labeled data but instead identify deviations from the norm.

In more complex monitoring scenarios, such as detecting sequential patterns over time, Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) models are ideal. These models are designed to detect anomalies in sequences of events, such as unusual login times or traffic spikes, by learning temporal dependencies.

Once the data is collected and threats are detected, the system must be capable of processing incoming data in real-time. Splunk is a powerful tool often used for this purpose, as it can index, search, and visualize large volumes of log data. By utilizing Splunk's data processing capabilities, large datasets are ingested, stored, and queried efficiently, enabling real-time threat detection. Splunk's indexing process is optimized to handle high-throughput data, making it an ideal solution for monitoring large networks and systems.

Splunk can ingest real-time data and use it to trigger alerts and execute responses. For instance, if a threshold is exceeded (e.g., traffic volume exceeds 1000 packets per second), Splunk can immediately trigger an alert. The traffic threshold can be mathematically represented as:

$$\text{AlertTriggered} = \text{“True”} \quad \text{if} \quad \text{Traffic Volume} > 1000\,\text{pps}$$

This allows for quick response actions, such as isolating a compromised system or blocking malicious IP addresses.

As the system processes more data, scalability becomes critical. Using a microservices architecture, the monitoring system can scale horizontally. In this architecture, each service (such as traffic analysis, threat detection, and alerting) can be scaled independently, allowing for optimized resource allocation. When the data load increases, new nodes can be added to maintain performance without overburdening the existing infrastructure.

Finally, as threats continuously evolve, the monitoring system must also evolve. Regular model retraining is vital for adapting to new attack patterns. The frequency of retraining can be determined by factors such as the appearance of new threat patterns and the incorporation of threat intelligence feeds. The retraining schedule can be adjusted dynamically:

$$\text{RetrainingFrequency} = f(\text{“New Data Patterns”}, \text{“Attack Intelligence”})$$

Additionally, feedback loops from security analysts or penetration testing teams help refine the system. For example, if the system flags false positives frequently, the decision threshold can be adjusted:

$$\text{Adjusted Threshold} = \text{Current Threshold} - \Delta$$

By continuously iterating and adapting the system based on new insights, the monitoring system remains effective at detecting new and unknown threats.

4. Conclusion

The article has demonstrated that achieving high efficiency in monitoring and anomaly detection systems in cybersecurity requires the implementation of distributed systems with high scalability, flexibility, and adaptability to new conditions. Real-time analysis of large datasets using Big Data technologies significantly enhances the ability of systems to detect anomalies and threats.

Developed mathematical models, such as scalability efficiency, availability assessment, and adaptability evaluation, enable the creation of more accurate predictions regarding the state of distributed systems and timely detection of potential threats.

References

1. Chen K. Fault detection, classification and location for transmission lines and distribution systems: a review on the methods / K. Chen, C. Huang, J. He. – High Voltage, 2016. – URL: https://doi.org/10.1049/hve.2016.0005.
2. Verdier G. Adaptive Mahalanobis Distance and k-Nearest Neighbor Rule for Fault Detection in Semiconductor Manufacturing / G. Verdier, A. Ferreira. – IEEE Transactions on Semiconductor Manufacturing, 2010. – URL: https://doi.org/10.1109/TSM.2010.2096538.
3. Tian J. Motor Bearing Fault Detection Using Spectral Kurtosis-Based Feature Extraction Coupled With K-Nearest Neighbor Distance Analysis / J. Tian, C. Morillo, M.H. Azarian, M. Pecht – IEEE Transactions on Industrial Electronics, 2015. – URL: https://doi.org/10.1109/TIE.2015.2478397.
4. Safizadeh M.S. Using multi-sensor data fusion for vibration fault diagnosis of rolling element bearings by accelerometer and load cell / M.S. Safizadeh, S.K. Latifi. – Information Fusion, 2013. – URL: https://doi.org/10.1016/j.inffus.2013.10.002.

Article: received 18.02.2024
revised 20.03.2024
printing adoption 10.04.2024

ІНТЕЛЕКТУАЛЬНИЙ МОНІТОРИНГ РОЗПОДІЛЕНИХ СИСТЕМ: ВИКОРИСТАННЯ МАШИННОГО НАВЧАННЯ ДЛЯ ВИЯВЛЕННЯ ТА АДАПТАЦІЇ ДО ЗМІНЮВАНИХ КІБЕРЗАГРОЗ

Р. Карпюк

Львівський національний університет імені Івана Франка,
вул. Університетська 1, Львів, 79000, Україна
e-mail: roman.karpiuk@lnu.edu.ua

Розглянемо важливість забезпечення стабільності, продуктивності та безпеки розподілених комп'ютерних систем шляхом інтеграції передових систем моніторингу, що використовують машинне навчання для виявлення аномалій у реальному часі та проактивної адаптації. Розглянуто проблеми обробки великих обсягів даних, архітектурної складності та зменшення хибних позитивних спрацьовувань, що створюють навантаження на аналітиків кібербезпеки. Основні компоненти охоплюють використання інструментів, таких як ELK або Splunk, для централізованого збору даних, застосування машинного навчання, наприклад, Random Forest, DensityFunction для виявлення відомих і невідомих загроз, а також використання масштабованих архітектур мікросервісів і еластичних хмарних платформ. Регулярне перенавчання моделей, динамічна корекція порогових значень і автоматичні сповіщення додатково поліпшують ефективність виявлення нових кіберзагроз.

Ключові слова: розподілені комп'ютерні системи, виявлення аномалій, кібербезпека, машинне навчання, моніторинг у реальному часі, хибні спрацьовування, масштабованість, архітектура мікросервісів, еластичні хмарні платформи, виявлення загроз, системні журнали, аналіз мережевого трафіка, системи виявлення вторгнень (IDS), перенавчання моделей, адаптивні системи моніторингу.