Application of Semantic Analysis and GMM Models for Anomaly Detection in Network Traffic

Maral B. Bekiyeva
Department of Applied Mathematics and Informatics Oguz han Engineering and Technology University of Turkmenistan
Ashgabat, Turkmenistan
successbmb@gmail.com

Gulshat O. Orazdurdyyeva
Department of Computer Sciences and Information Technologies Oguz han Engineering and Technology University of Turkmenistan
Ashgabat, Turkmenistan
gulshatorazdurdyyewa3@gmail.com

***Abstract*—Network traffic contains numerous patterns, and deviations from these patterns can indicate cyber-attacks or system failures. Traditional machine learning methods, such as the Gaussian Mixture Model (GMM), are effective in detecting anomalies, but do not provide meaningful interpretations of these anomalies. This paper presents an approach that integrates semantic analysis with GMM to improve anomaly detection accuracy and provide contextual insights into abnormal behavior in network traffic. Using cybersecurity ontologies and semantic reasoning, detected anomalies can be mapped to known cyber threats, improving the reliability of detection. The proposed method is evaluated using real-world network traffic logs, demonstrating its effectiveness in reducing false positives and enhancing interpretability.**

***Keywords*—Anomaly detection, network traffic analysis, Gaussian mixture model (GMM), semantic analysis, machine learning, cybersecurity, intrusion detection systems (IDS), unsupervised learning.**

I. Introduction

The rapid increase in cyberattacks poses a serious challenge to modern network security. Anomalous network traffic can indicate a variety of threats, including denial-of-service (DoS) attacks, unauthorized intrusions, and data exfiltration. Traditional methods for anomaly detection are based on statistical models and machine learning techniques, such as clustering and classification. Among these, the Gaussian Mixture Model (GMM) has proven to be effective in identifying outliers in network data. However, GMM has a significant drawback: it can identify anomalies but lacks interpretability, meaning that detected anomalies must be manually analyzed to determine their nature and relevance.

To address this limitation, semantic technologies can be leveraged. Semantic analysis provides a structured way to interpret and classify anomalies by linking them to ontologies, formal representations of knowledge that describe concepts, relationships, and rules. In the domain of cybersecurity, ontologies such as MITRE ATT&CK, STIX, and CybOX provide structured threat intelligence that can be used to categorize and explain network anomalies.

This paper proposes a hybrid approach that combines GMM-based anomaly detection with semantic reasoning, enabling automated classification and interpretation of detected anomalies. The primary contributions of this work are as follows.

1) Integration of semantic technologies and ontologies with machine learning for anomaly detection.

2) A methodology for mapping GMM-detected anomalies to known cyber threats using semantic reasoning.

3) An experimental evaluation that demonstrates the effectiveness of this approach in improving the precision and interpretability of network anomaly detection.

II. Related Work and Background

A. Gaussian Mixture Model (GMM) in Anomaly Detection

The Gaussian Mixture Model (GMM) is a probabilistic model that clusters data into multiple distributions, allowing for soft clustering – each data point can belong to multiple clusters with a certain probability. This makes GMM effective for detecting network anomalies because it can model complex traffic distributions and detect outliers in real-time.

B. Semantic Technologies and Ontologies in Cybersecurity

Cybersecurity Ontologies: Structured Knowledge of Threats

Ontologies provide structured knowledge bases that describe attack techniques, vulnerabilities, and network behaviors. Popular cybersecurity ontologies include:

- MITRE ATT&CK – A global framework categorizing cyber threats based on real-world attack techniques.
- STIX (Structured Threat Information Expression) – A language for sharing structured threat intelligence.

- CybOX (Cyber Observable Expression) – A format for describing network activities and anomalies.

How Semantic Analysis Improves Anomaly Detection

By combining GMM with semantic analysis, we can:

- Classify detected anomalies by mapping them to known cyber threats.
- Reduce false positives by filtering out normal variations.
- Describe the detected anomalies, making it easier for security analysts to respond.

Example: Suppose that GMM detects multiple failed login attempts followed by an unusual data transfer.

- Without semantic analysis: It is simply labeled as "anomalous".
- With semantic analysis: It is classified as a "Brute Force Attack" using the MITRE ATT&CK framework, triggering security recommendations.

Flowchart: Anomaly Detection with GMM and Semantic Technologies

The following diagram illustrates the step-by-step process of detecting anomalies using GMM combined with semantic technologies.

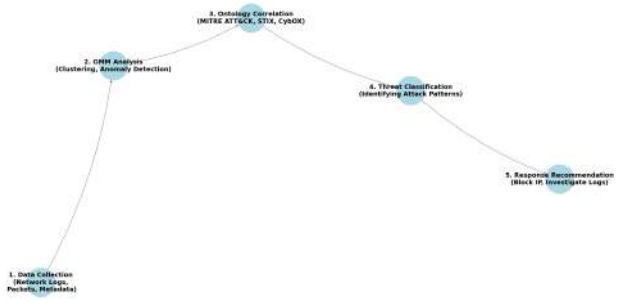

Figure 1. Detecting anomalies using GMM combined with semantic technologies.

Table I
Comparison of technologies for anomaly detection

Technology	Purpose	Advantages	Limitations
GMM	Detects network anomalies.	Flexible clustering, handles complex data.	No interpretability, high false positives.
MITRE ATT&CK	Classifies cyber threats.	Knowledge base of real-world attack methods.	Needs integration with detection systems.
STIX	Standardized cyber threat information sharing.	Improves collaboration across organizations.	Doesn't detect attacks on its own.
CybOX	Describes network events and attack indicators.	Helps analyze network traffic.	Requires integration with other tools.

By integrating GMM and semantic technologies, we create a smart, context-sensitive anomaly detection system that not only finds network anomalies but explains and classifies them for an effective cybersecurity response.

III. Proposed Methodology: Integrating GMM with Semantic Analysis

A. Overview of the Hybrid Approach

The core of our approach lies in the use of the Gaussian Mixture Model (GMM). Mathematically, the GMM is represented as a weighted sum of multiple Gaussian distributions:

The proposed system combines Gaussian Mixture Models (GMM) for anomaly detection with semantic reasoning to interpret and classify these anomalies. The hybrid approach leverages the strengths of both techniques to create a robust and scalable solution to detect and understand cybersecurity threats in network traffic. The system consists of three main components:

1) Anomaly detection using GMM: GMM is applied to network traffic data to identify clusters of normal and abnormal behaviors. This probabilistic model helps identify deviations from expected network patterns by grouping data points into distinct groups. Each group represents a specific pattern of behavior, and anomalies are detected when traffic deviates significantly from the norm.

2) Semantic Interpretation of Anomalies: The raw output from GMM, which identifies anomalies in network traffic, is mapped to a more meaningful context using cybersecurity ontologies. Ontologies provide structured frameworks that categorize different types of cybersecurity threats and network behaviors. This allows the system to interpret what each anomaly represents in terms of known attack types.

3) Threat Classification and Explanation: After identifying and interpreting the anomalies, the system proceeds to categorize them into specific types of cybersecurity threats (such as DDoS, phishing, or brute-force attacks). In addition, the system automatically generates textual explanations, helping security analysts understand the nature of the threat, its potential impact, and recommended countermeasures.

Together, these components enable the system to not only detect anomalies, but also contextualize and classify them, providing deeper insights into network security.

B. Ontology-Based Threat Mapping

Once GMM detects an anomaly, the next step is to determine the nature of the anomaly. This is done using an ontology-based reasoning system, which integrates the anomaly detection results with a comprehensive cybersecurity knowledge base. The reasoning system classifies the anomaly and provides an explanation leveraging the following two key components.

1. Semantic Knowledge Base: The knowledge base consists of a comprehensive set of cybersecurity-related

rules, patterns, and attack signatures. These patterns include both high-level attack types (such as DDoS, malware propagation, and brute-force attacks) and low-level network behaviors (such as packet frequency, failed login attempts, and unusual traffic patterns). The knowledge base is structured to accommodate a wide variety of network security events, providing context and detailed relationships between different types of anomaly and attack categories.

For example, the knowledge base might include the following.

- Behavioral patterns: Normal and abnormal behavior associated with different network protocols (e.g. HTTP, FTP, DNS) and network devices (e.g. routers, switches).
- Attack signatures: Known attack patterns such as port scanning, SQL injection, or Distributed Denial of Service (DDoS).
- Contextual rules: Rules that define the relationship between different network events, such as multiple failed login attempts leading to a brute-force attack or a sudden spike in traffic indicating a potential DDoS attack.

The knowledge base is regularly updated to include the latest attack techniques and evolving network traffic patterns, ensuring the reasoning system remains effective over time.

2. Reasoning Engine: The reasoning engine is responsible for inferring the most likely type of attack based on the data provided by the GMM and the knowledge stored in the semantic knowledge base. The engine applies semantic reasoning techniques to map the anomalies detected by GMM to specific attack types. The reasoning process follows a rule-based inference mechanism, which can be implemented using logical rules or machine learning models.

The engine operates by processing the output from GMM, which includes a set of anomaly scores or data points. Then these are compared against the patterns and relationships in the knowledge base. The system uses inference rules (e.g., if X and Y occur simultaneously, the event is classified as Z) to determine the most likely attack type. The reasoning engine can also handle complex scenarios, in which multiple anomalies must be considered together to accurately classify an attack type.

For example, if the system detects an anomaly involving unusually high packet frequency along with a high number of failed authentication attempts, the reasoning engine might infer that this is a brute-force attack. Maps these two factors (high packet frequency and authentication failures) to predefined rules in the knowledge base that describe brute-force attacks.

C. Benefits of the Hybrid Approach

By combining GMM and semantic reasoning, the proposed system offers several key advantages over traditional methods.

1) Scalability: GMM is highly scalable and can handle large volumes of network traffic data, making it suitable for real-time network monitoring.

2) Accurate Threat Classification: The integration of semantic reasoning ensures that detected anomalies are accurately classified into meaningful cybersecurity threat categories. This reduces false positives and improves the reliability of the system.

3) Explainability: The system generates automated explanations of detected threats, providing security analysts with clear and actionable insights into the nature of the attack, its potential impact, and appropriate countermeasures.

4) Adaptability: The system's knowledge base can be updated with new attacks and patterns, which allows it to remain relevant in the face of evolving cyber threats.

IV. Experimental Evaluation

In this section, we evaluate the effectiveness of the proposed methodology by conducting experiments using real-world network datasets. The evaluation focuses on evaluating the performance of the anomaly detection system, its ability to classify network traffic anomalies, and its overall accuracy in identifying various types of cyber threats.

A. Dataset and Preprocessing

The proposed method is evaluated using two widely recognized datasets commonly used in network intrusion detection research. These data sets provide diverse and comprehensive examples of both traditional and modern attack scenarios, ensuring that the proposed system is tested under varied conditions.

- KDD Cup 1999: This data set is one of the most well known in the field of intrusion detection and contains network traffic data captured from a simulated military network. It includes both normal traffic and multiple types of attacks, such as DoS (Denial of Service), U2R (User to Root), R2L (Remote to Local), and probing attacks. The data set is used to assess the ability of the system to detect different types of attacks and to assess its general performance in intrusion detection.
- CICIDS2017: This data set contains modern attack scenarios, including advanced threats such as botnets, DoS attacks, and malware activities. It offers more realistic network traffic compared to KDD Cup 1999, including a mix of benign and malicious traffic from both known and unknown attack patterns. The CICIDS2017 data set is designed to test the system's adaptability to more contemporary attack vectors.

Data Preprocessing: The raw data from these datasets are preprocessed to extract key features that are relevant for the detection and classification of anomalies. The preprocessing steps include the following.

- Packet Sizes and Intervals: The size of each packet and the time intervals between packets are important features in detecting anomalies. For example, in a DDoS attack, there might be a sudden surge in packet sizes or a high frequency of packet transmissions in a short period.

- Source and Destination IP Addresses: The source and destination IP addresses help to identify the origin and target of the network traffic. Suspicious behavior, such as traffic from a single IP address targeting multiple destinations, may indicate a botnet or other malicious activity.
- Protocol Types and Connection Attempts: The types of protocols (e.g., HTTP, FTP, ICMP) and the number of connection attempts are key indicators of malicious activity. Abnormal patterns, such as multiple failed connection attempts using a particular protocol, may point to a brute-force attack or scanning attempts.

By extracting these features, the data become suitable for analysis by the GMM-based anomaly detection system, enabling the identification of deviations from normal behavior in the network traffic.

B. Performance Metrics

The effectiveness of the proposed anomaly detection method is evaluated using several standard performance metrics, which provide insight into the system's ability to detect anomalies and classify threats accurately. These metrics include:

1) Accuracy: Accuracy is one of the most straightforward metrics used to evaluate the overall performance of the detection system. It is calculated as the ratio of correctly identified anomalies (both true positives and true negatives) to the total number of instances. High accuracy indicates that the system is good at distinguishing between normal and anomalous traffic.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where:

- TP: True Positives (correctly identified anomalies)
- TN: True Negatives (correctly identified normal traffic)
- FP: False Positives (normal traffic incorrectly identified as anomalous)
- FN: False Negatives (anomalous traffic missed by the system)

2) Precision: Precision measures the accuracy of positive predictions made by the system. In the context of anomaly detection, precision represents the proportion of correctly detected anomalies out of all instances that were classified as anomalous by the system. A high precision value indicates that the system produces few false positives.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Precision is particularly important in situations where false positives are costly, such as in a security system where falsely flagging normal traffic as anomalous may disrupt business operations or waste resources.

3) Recall: Recall (also known as Sensitivity or True Positive Rate) measures the ability of the system to detect all actual anomalies. It is the proportion of true anomalies that were correctly identified by the system. A high recall value indicates that the system is good at identifying most of the malicious activity present in the network.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

4) False Positive Rate (FPR): The False Positive Rate measures the proportion of normal traffic that is incorrectly classified as anomalous. It is calculated as the number of false positives divided by the total number of actual normal instances. A low FPR is crucial for ensuring that the system does not generate too many false alarms, which can lead to alert fatigue among security analysts.

$$FPR = \frac{FP}{FP + TN} \tag{4}$$

A high FPR can reduce the effectiveness of the system, as analysts may ignore or overlook legitimate alerts if too many false positives are raised.

5) F1 score: The F1 score is the harmonic mean of precision and recall. It provides a balanced measure of the system's performance when both false positives and false negatives are important to consider. The F1 score is particularly useful when there is an imbalance between the number of normal and anomalous traffic instances.

$$F1 = 2 \times \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \right) \tag{5}$$

Using these metrics, we can comprehensively evaluate the performance of the proposed anomaly detection system, ensuring that it strikes an optimal balance between identifying threats and minimizing false alarms. Furthermore, performance can be compared between different datasets and compared against existing anomaly detection methods to demonstrate the advantages of the proposed approach.

C. Results and Analysis

The experimental evaluation reveals several key insights into the performance of the proposed hybrid system, which combines Gaussian mixture models (GMM) with semantic analysis for network anomaly detection and threat classification. The results were analyzed in comparison to using GMM alone, as well as against existing baseline methods.

1) Performance of GMM Alone: When applied independently, GMM achieves high anomaly detection rates, demonstrating its effectiveness in detecting deviations from normal network behavior. The model is able to identify various outliers in traffic data and can effectively group normal and anomalous behavior into separate

groups. However, while GMM excels at detecting anomalies, it suffers from a high false positive rate (FPR). This means that benign traffic, which exhibits slight deviations due to network variability, is often misclassified as anomalous. As a result, security analysts may be overwhelmed by a large number of false alarms, making it difficult to prioritize real threats.

Despite this drawback, the strength of GMM lies in its unsupervised nature, which allows it to work with diverse and large datasets without requiring labeled training data. This makes it particularly useful in real-world scenarios where labeled data may be scarce or hard to obtain.

2) Improvement with Semantic Analysis: When combined with semantic analysis and the use of a cybersecurity ontology, the system shows a substantial improvement in both the accuracy and reliability of anomaly detection:

- Improved Classification Accuracy: Using the semantic analysis component, the system can map detected anomalies to known cyber threat categories, such as DDoS, brute-force attacks, and malware. This contextualization improves the system's ability to classify anomalies accurately. For example, a high number of failed login attempts combined with abnormal packet patterns can now be confidently classified as a brute-force attack based on the reasoning engine's inference from the ontology. This mapping process significantly improves the overall classification accuracy, as the system now classifies detected anomalies with a clear understanding of their underlying cybersecurity implications.
- Reduced False Positives: One of the primary advantages of integrating semantic reasoning is the reduction in false positives. The semantic engine helps filter out benign anomalies that may have been incorrectly flagged by GMM. For instance, certain traffic spikes or slight variations in packet sizes that are normal in specific network contexts can be recognized as nonthreatening based on the semantic rules in the ontology. This selective filtering minimizes the risk that benign traffic is misclassified as an attack, allowing security teams to focus on genuine threats. Consequently, the false positive rate (FPR) is significantly reduced, leading to a more efficient and manageable workflow for cybersecurity professionals.
- Automated Threat Explanations: Another key benefit of integrating semantic analysis is the automatic generation of threat explanations. When an anomaly is detected and classified, the system provides detailed information on the nature of the detected threat. These explanations include the type of attack, the key features or behaviors that led to the classification, and potential mitigation strategies. For example, in the case of a DDoS attack, the explanation might describe the abnormal traffic patterns observed, such as the volume of incoming requests and the specific targets affected. This transparency helps cybersecurity analysts make more informed decisions, reducing the time required to understand the nature of the attack and respond effectively.

3) Comparative Performance: To further validate the effectiveness of the proposed hybrid approach, a comparative analysis was performed against other existing methods in the field of anomaly detection.

- Accuracy: The hybrid system showed a significant improvement in accuracy over GMM alone, demonstrating a better balance between true positive detections and minimizing false positives. When tested on both the KDD Cup 1999 and CICIDS2017 datasets, the hybrid system outperformed traditional methods, particularly in detecting complex attack patterns that GMM alone struggled to identify.
- Precision and Recall: Precision and recall were also improved in the hybrid model. The integration of semantic reasoning allowed the system to be more selective in flagging anomalies, leading to higher precision in attack classification. At the same time, the system recall rate remained high, ensuring that most attacks were still detected. This balance is critical in ensuring that the system does not sacrifice the detection rate for fewer false alarms, a common issue in many anomaly detection systems.
- False Positive Rate (FPR): As mentioned, the false positive rate saw a significant decrease when semantic analysis was applied. This is crucial for operational efficiency, as high FPR can lead to alert fatigue, where security teams become desensitized to the large number of false alarms. By reducing FPR, the system ensures that security personnel can focus their efforts on investigating real threats.

4) Performance of GMM Alone: The results of the experimental evaluation suggest that the proposed hybrid approach offers a highly practical solution for real-time network traffic monitoring and cyber threat detection.

- Enhanced Threat Detection: By combining the statistical clustering of GMM with the semantic understanding provided by the ontology, the system can detect and classify a wide range of cyber threats more accurately and reliably than GMM alone.
- Operational Efficiency: The reduction in false positives and the ability to provide automated and understandable explanations of detected threats help improve the overall operational efficiency of cybersecurity teams. Analysts can make quicker decisions, reduce response times, and allocate resources more effectively to mitigate real threats.
- Adaptability: The system's reliance on an evolving semantic knowledge base means that it can be continuously updated with new attack patterns and emerging threats, making it adaptable to the changing landscape of cyber attacks.

5) Future Work: Although the current results are promising, further improvements can be made to increase the robustness and efficiency of the system.

- Expansion of the Knowledge Base: The knowledge base can be expanded to include additional attack patterns, protocols, and more fine-grained network behaviors. This would improve the system's ability to detect new and evolving threats.
- Real-Time Deployment: Future work will focus on optimizing the system for real-time deployment in live networks, ensuring that it can scale to handle large volumes of traffic without compromising detection performance.
- Integration with Other Security Tools: The system could also be integrated with other security solutions, such as firewalls and intrusion prevention systems (IPS), to provide a more comprehensive security infrastructure.

D. System Diagram

A high-level overview of the proposed hybrid system architecture can be visualized in the form of a system diagram. This diagram will help illustrate the components involved in the anomaly detection pipeline, showing how the Gaussian Mixture Model (GMM) interacts with the semantic analysis component to improve detection accuracy and reduce false positives.

The system consists of the following key components:

1) Data Collection: Raw network traffic data data collected and preprocessed. This data includes packet sizes, IP addresses, protocol types, and connection attempts.

2) Anomaly Detection (GMM): The Gaussian Mixture Model (GMM) is used to model network traffic and identify potential anomalies based on statistical deviations.

3) Semantic Analysis: Once anomalies are detected, the system uses an ontology-based semantic reasoning engine to interpret and classify the anomalies. This engine links the detected anomalies to known cybersecurity threats.

4) Threat Classification: The system classifies the anomalies into specific cyber threats such as DoS, DDoS, brute-force attacks, and more.

5) Explanations and Decision Support: The system generates automated explanations for detected threats, providing security analysts with context and reasoning behind each classification.

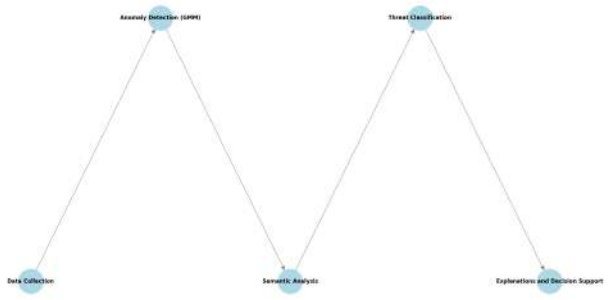

Figure 2. System diagram which shows the flow of data from one component to another.

E. Explanation of the System Diagram

- Data Collection: This block represents the initial step where raw network traffic is gathered and preprocessed. Preprocessing includes feature extraction such as packet sizes, IP addresses, protocols, and connection attempts. The preprocessed data are then passed to the next stage for anomaly detection.
- Anomaly Detection (GMM): GMM models normal traffic behavior and flags deviations as anomalies. These are passed to semantic analysis.
- Semantic analysis: Using a cybersecurity ontology, the system interprets anomalies by matching them to known attack patterns (e.g., DoS, brute force, botnets).
- Threat Classification: The mapped anomalies are categorized (DDoS, malware, etc.) to assess the type and potential impact.
- Explanations and Decision Support: The system generates human-readable explanations to help analysts understand and respond to threats effectively.

V. Conclusion

This paper presents a novel hybrid approach for network anomaly detection and threat classification that combines Gaussian Mixture Model (GMM)-based anomaly detection with semantic reasoning. The integration of these two techniques significantly improves the effectiveness of cybersecurity analysis by addressing some of the key limitations of traditional machine learning approaches.

References

[1] A. Lakhina, M. Crovella and Diot, C. “Diagnosing network-wide traffic anomalies.” ACM SIGCOMM Computer Communication Review, 34(4), 2004, pp. 219–230.

[2] B. Varghese and R. Buyya “Next Generation Cloud Computing: New Trends and Research Directions”. Elsevier. 2018.

[3] C. M. Bishop “Pattern Recognition and Machine Learning”. Springer. 2006.

[4] C. Krügel, D. Mutz, W. Robertson and G. Vigna “Bayesian event classification for intrusion detection.” Proceedings of the 19th Annual Computer Security Applications Conference, 2003, pp. 1–9.

[5] G. Ahmed, A. N. Mahmood and J. Hu “A survey of network anomaly detection techniques.” Journal of Network and Computer Applications, 60, 2016, pp. 19–31.

[6] G. O. Orazdurdyyeva “Simulating network conditions and DDoS attack scenarios using NS-3 technology”, EDA Conference, 2024.

[7] J. Ren, X. Lyu, Y. Zhang, and H. Wen “Anomaly Detection for Network Traffic Using Deep Autoencoder Gaussian Mixture Model.” IEEE International Conference on Communications, 2019, 1–6.

[8] M. B. Bekiyeva “Modeling network traffic dynamics under DDoS attacks using differential equations”, EDA Conference, 2024.

[9] M. A. Pimentel, D. A. Clifton, L. Clifton and L. Tarassenko “A review of novelty detection.” Signal Processing, 99, 2014, pp. 215–249.

[10] V. Chandola, A. Banerjee and V. Kumar “Anomaly detection: A survey”. ACM Computing Surveys (CSUR), 41(3), 2009, 1–58. DOI: 10.1145/1541880.1541882.

ПРИМЕНЕНИЕ СЕМАНТИЧЕСКОГО АНАЛИЗА И МОДЕЛЕЙ GMM ДЛЯ ОБНАРУЖЕНИЯ АНОМАЛИЙ В СЕТЕВОМ ТРАФИКЕ

Бекиева М. Б., Ораздурдыева Г. О.

Сетевой трафик содержит многочисленные шаблоны, и отклонения от этих шаблонов могут указывать на кибератаки или сбои системы. Традиционные методы машинного обучения, такие как модель гауссовской смеси (GMM), эффективны для обнаружения аномалий, но не дают содержательной интерпретации этих аномалий. В этой статье представлен подход, который объединяет семантический анализ с GMM для повышения точности обнаружения аномалий и предоставления контекстуальных сведений об аномальном поведении в сетевом трафике. Используя онтологии кибербезопасности и семантическое обоснование, обнаруженные аномалии можно сопоставить с известными киберугрозами, что повышает надежность обнаружения. Предлагаемый метод оценивается с использованием журналов реального сетевого трафика, демонстрируя его эффективность в снижении ложных срабатываний и улучшении интерпретируемости.

Received 23.03.2025