Abstract

Anomaly detection in network traffic is a vital task for ensuring network security and performance. It aims to identify abnormal or malicious traffic patterns that deviate from normal behavior. Various techniques have been proposed for this task, such as statistical analysis, rule-based detection, and machine learning. Machine learning is a powerful and flexible approach that can learn from data and adapt to changing network conditions. However, machine learning also faces some challenges, such as feature selection, data imbalance, scalability, and interpretability.

This paper reviews some of the recent advances and applications of machine learning for anomaly detection in network traffic, focusing on four main aspects: (1) data preprocessing and feature engineering, (2) supervised, and unsupervised learning methods, (3) deep learning and neural networks, and (4) evaluation metrics and benchmarks. The paper also discusses some of the open issues and future directions for this research area.

Data preprocessing and feature engineering are essential steps for preparing the network traffic data for machine learning models. They involve cleaning, filtering, aggregating, transforming, and selecting the relevant features that can capture the characteristics of normal and abnormal traffic. Some of the common features used for anomaly detection are flow-based features, such as source and destination IP addresses, ports, protocols, packet sizes, durations, inter-arrival times, etc., and payload-based features, such as content types, keywords, signatures, etc.

This paper provides a comprehensive overview of machine learning approaches to anomaly detection in network traffic. It also identifies some of the challenges and opportunities for future research in this domain.

Introduction

Nowadays, the opportunities for hacking methods are large-scale. Whenever surfing on the internet we face the issue with the network, moreover anomalies in network traffic. Problems with the network traffic might be caused by external factors, known as physical troubles in network devices, and related them the connectedness conductors. Internal issues with network traffic are crucial since the physical issues are thought to be quickly rectified. For deviation in network traffic had to face the incidents, such as Port scans, DDoS, and Web attacks. Port scans relied on the purpose of hackers, and what they want to gain. Based on the scanning of the open ports through which TCP and UDP ports are able to receive packets, each and everyone would be able to capture the packets between endpoints of the network flow. A DDoS (Distributed Denial-of-Service) attack is rare, although due to this attack, there is a crucial loss of workflow time that is caused by the attack's impacts on account of large amounts of network traffic packets. Exactly, by exploring the network traffic problems we used the program of capturing the packets in the network Wireshark and its required files for the practical part of using Machine Learning.

Main part

1. Anomaly detection using machine learning

Anomaly detection in network traffic is a technique to enhance network security by identifying abnormal or malicious traffic patterns that deviate from normal behavior. It can help prevent or mitigate the impact of cyberattacks, such as port scans, DDoS, and web attacks. Machine learning is a flexible and powerful approach that can learn from data and adapt to changing network conditions. It can also overcome some of the limitations of traditional methods, such as statistical analysis and rule-based detection, which may require prior knowledge, manual tuning, or frequent updates.

Machine learning for anomaly detection in network traffic involves several steps, such as data preprocessing, feature engineering, model selection, training, testing, and evaluation. Data preprocessing and feature engineering are essential for preparing the network traffic data for machine learning models. They involve cleaning, filtering, aggregating, transforming, and selecting the relevant features that can capture the characteristics of normal and abnormal traffic. Some of the common features used are flow-based features, such as source and destination IP addresses, ports, protocols, packet sizes, durations, inter-arrival times, etc., and payload-based features, such as content types, keywords, signatures, etc.

Model selection is the process of choosing a suitable machine learning algorithm for the anomaly detection task. There are various types of machine learning algorithms, such as supervised, semi-supervised, unsupervised, and deep learning. Supervised learning algorithms require labeled data for training and testing. They can perform binary or multiclass classification of network traffic anomalies. Some examples are stochastic gradient descent (SGD), support vector machines (SVM), k-nearest neighbor (K-NN), Gaussian naive Bayes (GNB), decision tree random forest (RF), and AdaBoost (AB). Unsupervised learning algorithms do not require labeled data for training and testing. They can perform clustering or outlier detection of network traffic anomalies. Some examples are k-means clustering (KMC), density-based spatial clustering of applications with noise (DBSCAN), local outlier factor (LOF), isolation forest (IF), and one-class SVM (OCSVM). Deep learning algorithms are a subset of machine learning algorithms that use multiple layers of artificial neural networks to learn complex patterns from data. They can perform various tasks such as feature extraction, dimensionality reduction, classification, or regression of network traffic anomalies. Some examples are autoencoders (AE), convolutional neural networks (CNN), recurrent neural networks (RNN), long short-term memory (LSTM), and generative adversarial networks (GAN).

2. Preprocessing

Preprocessing is an important step for anomaly detection in network traffic using machine learning. It involves preparing the network traffic data for machine learning models by cleaning, filtering, aggregating, transforming, and selecting the relevant features that can capture the characteristics of normal and abnormal traffic. Preprocessing can improve the accuracy and capability of anomaly detection systems by reducing data dimensionality, noise, and redundancy.

Some of the common preprocessing techniques for anomaly detection in network traffic are:

Data cleaning: removing or correcting erroneous, incomplete, or inconsistent data records.

Data filtering: selecting a subset of data records that are relevant for the analysis task, such as removing background traffic or focusing on a specific protocol or service.

Data aggregation: combining multiple data records into a single record by applying summary statistics, such as mean, median, standard deviation, etc.

Data transformation: applying mathematical or logical operations to change the scale, format, or distribution of the data values, such as normalization, standardization, logarithm, etc.

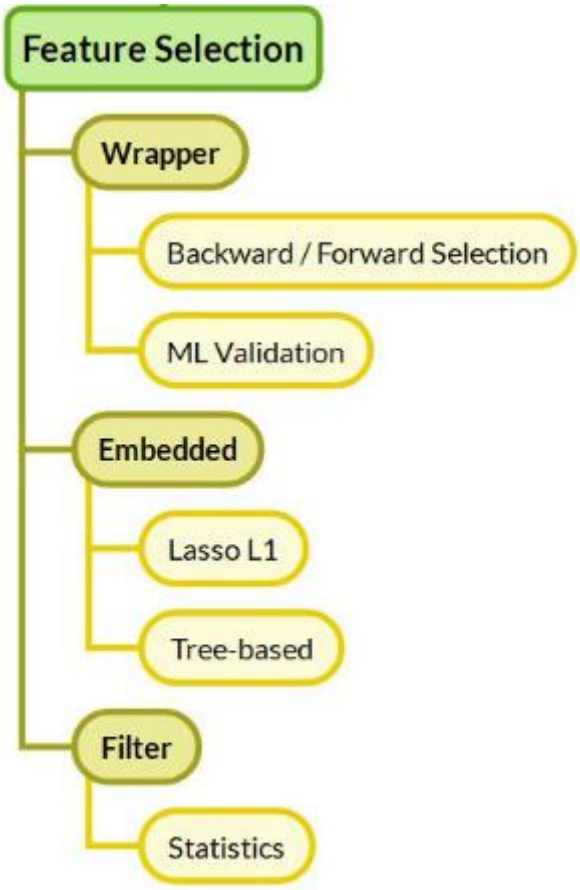
Feature construction: deriving new features from the raw data or existing features by applying domain knowledge or heuristic rules, such as extracting payload content, computing time-based statistics, etc.

Feature selection: finding the most relevant subset of features from a large candidate set by applying automated methods, such as filter methods, wrapper methods, or embedded methods.

Preprocessing can also be tailored to specific types of anomalies or attacks, such as port scans, DDoS attacks, web attacks, etc., by using targeted content parsing or context-sensitive features. Preprocessing can also leverage In-band Network Telemetry (INT)

3. Feature selection

The finding of anomalies is coming from the first actions with choosing the headings. Selecting features act a significant and concentrated part in continuing the process of detecting deviations in network traffic. It helps carry over from preprocessing to filtering the attacks. But a set collection of influential features estop on some nuances to collect full aspects. According to Zimek and other authors from taken article "Analysis of network traffic features for anomaly detection" which is about the problems of classification and implementation of machine learning, found in the part of "Problems of high-dimensionality for classification" that the issue relied on the dimensionality of features, and has a strong effect on some classification techniques, yet in clustering and anomaly detection in general. The feature selection which is dependent on account of the high amount of instances often faces classification issues and to ease them, must be useful methods in the future, such as the filter method, wrapped method, and embedded method as shown in **picture 1**, and as reported by article "Automated Feature Selection for Anomaly Detection in Network Traffic Data" in the section about feature selection classes and their methods.



Picture 1. Feature selection filters

For mandatory stuff, filter selection was needed, but for complete information, this article "An introduction to variable and feature selection" about these types can be helpful. Additionally, for collecting the features might be used importance of each from whole features to compare them and select one set of features for dermining the attack. Pursuant to Kostas's methods he mentioned that feature selection should be identified by its weight significance, and Random Forest would be the best way to define it.

4. Anomaly detection algorithms and model

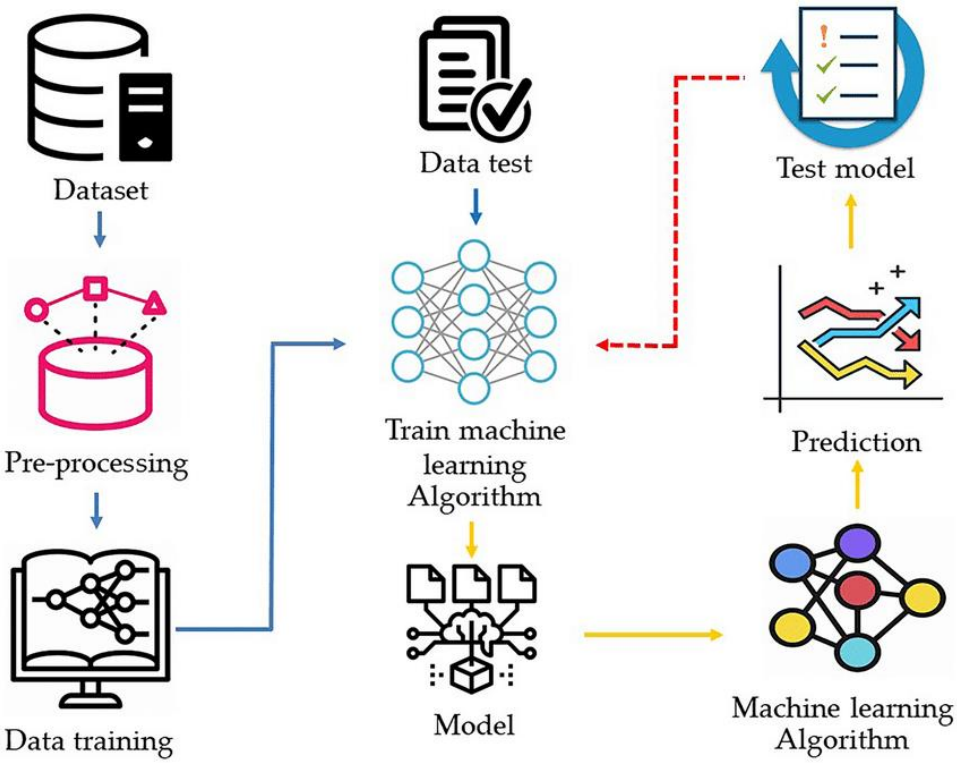
This section about algorithms and models touched on the themes to detect anomalies automatically and to improve the process of prediction. After pre-processing and feature selection processes in the operation of machine learning algorithms are needed. The machine learning algorithms originate from unsupervised and supervised types and are consigned to predict accurately. The supervised type of predicting algorithm was the most convenient, because of the focus on the rapidity of solving the issue with network anomalies. Supervised learning has algorithms with each peculiarity on it. Basically, supervised learning is divided into two ways of algorithms' direct:

- Regression
- Classification

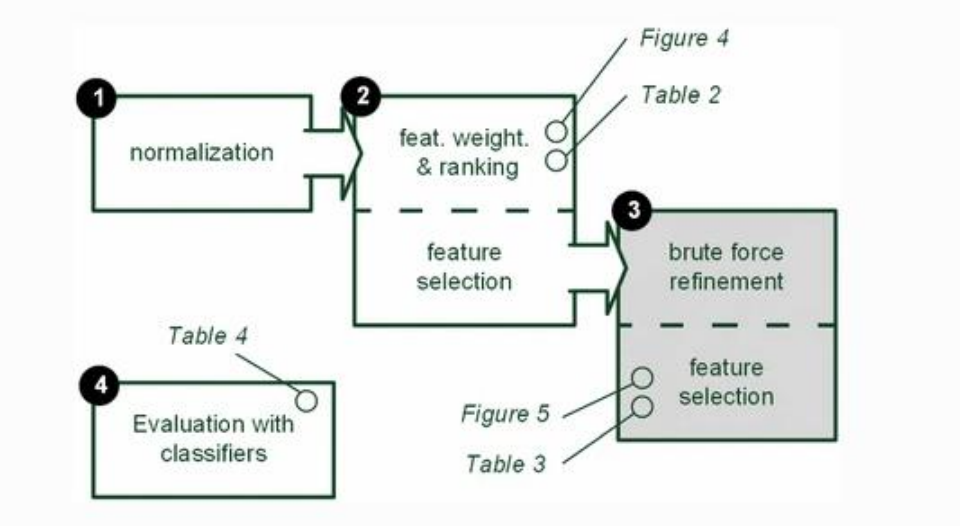
In classification process of machine learning has KNN, Naive Bayes, SVM, and Logistic Regression which is in the process of regression too. Furthermore, machine learning has other algorithms such as Random Forest, Decision Tree, Iterative Dichotomiser 3 algorithm,

Adaboost algorithm, and Multilayer Perception along with others are able to find in the article about types of machine learning algorithms. For awareness required to take a look at definitions of each type of algorithm for the practice part. The first instance is KNN, known as the k-nearest neighbors algorithm identified as non-parametric and comes from supervised learning to predict and make the classification processes. Next is the Random Forest algorithm derive from an ensembling type of machine learning needed for classification and regression operations consisting of organized decision trees at the time of training. The AdaBoost algorithm too emerges as an ensembling method of machine learning and interprets it as a boosting approach. The fourth is Iterative Dichotomiser 3 specified short name ID3 refers to the algorithm which is building a decision tree and was invented by Ross Quinlan.

The models are constructed by algorithms and other staged operations such as pre-processing, feature selection, and prediction. But in fact using machine learning, first must be taken dataset. The dataset might be pre-processed to be normalized. The next step should be a split to train and test data by selecting the features and target to predict. After dividing the type algorithm should be chosen. The penultimate stage is in predicting and the last is covered in evaluating the forecasting accuracy of the prediction model which has shown in **picture 2**.



Picture 2. Machine learning model



Picture 3. The prediction model

In **picture 3** from the article analysis of network traffic features for anomaly detection was likewise valuable for improving the model in the detection of anomalies machine learning.

Methodology

In this methodology section, we decided to consider Kostas explored research about anomalies in network traffic, especially from his article on the CICIDS2017 dataset which includes the most common network traffic attacks and commends the real-world data Wireshark recorded pcap in CSV format files. First, we should filter the attacks, and how many they are by scanning the dataset. By the result of scanning the score of attacks, we are able to see how many they are. If we consider them:

- DDoS: 41 835
- Port Scan attacks: 158 930
- Web attacks: 21

Second, we took these files with common attacks such as Web attacks, DDoS, and Port scans. In the whole files were similar features with a high dimensional amount. All files contain 86 recorded features, such as Source, Destination IP addresses and ports Total Forward and Backward packets, and others. In the feature selection stage was identified 5 relevant features were to provide the training and test process. But for each attack selected particular features, comparatively as for DDoS were selected 5 peculiar features:

- Bwd Packet Length Std
- Total Backward Packets
- Fwd IAT Total
- Flow Duration
- Flow IAT Min

Especially for the Web attack other 5 features:

- Bwd Packet Length Std

- Total Length of Fwd Packets
- Flow Bytes/s
- Flow IAT Max
- Bwd Packet Length Max

Next attack Port scan demanded in 5 various features:

- Flow Bytes/s
- Total Length of Fwd Packets
- Fwd IAT Total
- Flow Duration
- Fwd Packet Length Max

The next step was the last step in determining the most favorable algorithms and it was urgent for machine learning implementation. We examined 7 algorithms, called Random Forest, ID3 (Iterative Dichotomiser 3), Adaboost algorithm, and Multilayer Perception, KNN (K-nearest neighbors), and evaluated each of their prediction models. After examination of algorithms revealed 4 of them presented as the best ones. A list of accurate algorithms was:

- Adaboost algorithm
- Random Forest
- ID3
- K-nearest neighbors

The high accuracy of the four algorithms was greater than 90 percentage of predicting.

Discussion

Anomaly detection in network traffic using machine learning is a challenging and important research area that has many applications for network security and performance. Machine learning can provide powerful and flexible methods to learn from data and adapt to changing network conditions. However, machine learning also faces some challenges, such as data preprocessing, feature selection, model selection, parameter tuning, evaluation metrics, and interpretability.

Data preprocessing is a crucial step for preparing the network traffic data for machine learning models. It involves cleaning, filtering, aggregating, transforming, and selecting the relevant features that can capture the characteristics of normal and abnormal traffic. Data preprocessing can improve the accuracy and capability of anomaly detection systems by reducing data dimensionality, noise, and redundancy.

Feature selection is another important step for finding the most relevant subset of features from a large candidate set. Feature selection can enhance the performance of machine learning models by removing irrelevant or redundant features, reducing computational complexity, and avoiding overfitting. Feature selection can be done by applying automated methods, such as filter methods, wrapper methods, or embedded methods.

Conclusion

In conclusion, based on the results of our methods dealing with anomalies, Web attacks, DDoS, Port Scan attacks in network traffic, demonstrated that it can be crucial for being safe.

We explored only 3 attacks and in the future would be more than we have dealt with in the beginning.

References

Ahmed, T., Oreshkin, B., & Coates, M. (2007, April). Machine learning approaches to network anomaly detection. In *Proceedings of the 2nd USENIX workshop on Tackling computer systems problems with machine learning techniques* (pp. 1-6). USENIX Association. https://www.usenix.org/legacy/event/sysml07/tech/full_papers/ahmed/ahmed.pdf?ref=driverlayer.com/web

Alomari, O., & Othman, Z. A. (2012). Bees algorithm for feature selection in network anomaly detection. *Journal of applied sciences research*, *8*(3), 1748-1756. https://www.researchgate.net/profile/Osama-Alomari/publication/267427043_Bees_Algorithm_for_feature_selection_in_Network_Anomaly_detection/links/54b2b22d0cf28ebe92e2d901/Bees-Algorithm-for-feature-selection-in-Network-Anomaly-detection.pdf

Amrollahi, M., Hadayeghparast, S., Karimipour, H., Derakhshan, F., & Srivastava, G. (2020). Enhancing network security via machine learning: opportunities and challenges. *Handbook of big data privacy*, 165-189 doi: 10.1007/978-3-030-38557-6_8

Asmuss, J., & Lauks, G. (2015, August). Network traffic classification for anomaly detection fuzzy clustering based approach. In *2015 12th International conference on fuzzy systems and knowledge discovery (FSKD)* (pp. 313-318). IEEE. doi: 10.1109/FSKD.2015.7381960

Bhuyan, M. H., Bhattacharyya, D. K., & Kalita, J. K. (2013). Network anomaly detection: methods, systems and tools. *Ieee communications surveys & tutorials*, *16*(1), 303-336. doi: 10.1109/SURV.2013.052213.00046

Chen, S., Huang, Z., Zuo, Z., & Guo, X. (2016, October). A feature selection method for anomaly detection based on improved genetic algorithm. In *2016 4th International Conference on Mechanical Materials and Manufacturing Engineering* (pp. 186-189). Atlantis Press. https://www.atlantis-press.com/proceedings/mmme-16/25859790

Iglesias, F., & Zseby, T. (2015). Analysis of network traffic features for anomaly detection. *Machine Learning*, *101*, 59-84. https://link.springer.com/article/10.1007/s10994-014-5473-9#Sec23

Imamverdiyev, Y., & Sukhostat, L. (2016, October). Anomaly detection in network traffic using extreme learning machine. In 2016 IEEE 10th international conference on application of information and communication technologies (AICT) (pp. 1-4). IEEE doi: 10.1109/ICAICT.2016.7991732

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), 1157-1182. https://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf?ref=driverlayer.com/web

Kostas K. (2021, June). Anomaly detection in Networks Using Machine Learning. *Computer Science and Electronic Engineering - CSEE*. https://github.com/kahramankostas/Anomaly-Detection-in-Networks-Using-Machine-Learning/blob/master/README.md

Kostas K. (2018, August). Anomaly Detection in Networks Using Machine Learning. *Heriot-Watt University* (pp. 10-51) https://www.researchgate.net/publication/328512658_Anomaly_Detection_in_Networks_Using_Machine_Learning

Kumari, R., Singh, M. K., Jha, R., & Singh, N. K. (2016, March). Anomaly detection in network traffic using K-mean clustering. In *2016 3rd international conference on recent advances in information technology (RAIT)* (pp. 387-393). IEEE. doi: 10.1109/RAIT.2016.7507933

Kurniawan, D. A., Wibirama, S., & Setiawan, N. A. (2016, October). Real-time traffic classification with Twitter data mining. In *2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE)* (pp. 1-5). IEEE. doi: 10.1109/ICITEED.2016.7863251

Münz, G., Li, S., & Carle, G. (2007, September). Traffic anomaly detection using k-means clustering. In *Gi/itg workshop mmbnet* (Vol. 7, No. 9). https://www.net.in.tum.de/projects/dfg-lupus/files/muenz07k-means.pdf

Nakashima, M., Sim, A., Kim, Y., Kim, J., & Kim, J. (2021). Automated feature selection for anomaly detection in network traffic data. *ACM Transactions on Management Information Systems (TMIS)*, *12*(3), 1-28. https://doi.org/10.1145/3446636

Patgiri, R., Varshney, U., Akutota, T., & Kunde, R. (2018, November). An investigation on intrusion detection system using machine learning. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1684-1691). IEEE. doi: 10.1109/SSCI.2018.8628676

Ray, S. (2019, February). A quick review of machine learning algorithms. In *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)* (pp. 35-39). IEEE. https://books.google.kz/books?hl=ru&lr=&id=XAqhDwAAQBAJ&oi=fnd&pg=PA19&dq=machine+learning+algorithms+all+types&ots=r2No7XFeQm&sig=5HJcJY7_DRWwrULDJefcPtnNrj8&redir_esc=y#v=onepage&q=machine%20learning%20algorithms%20all%20types&f=false

Ringberg, H., Soule, A., Rexford, J., & Diot, C. (2007, June). Sensitivity of PCA for traffic anomaly detection. In *Proceedings of the 2007 ACM SIGMETRICS international conference on Measurement and modeling of computer systems* (pp. 109-120). https://doi.org/10.1145/1254882.1254895

Rojas J. S. (2018). IP Network Traffic dataset. *Kaggle*. https://www.kaggle.com/datasets/jsrojas/ip-network-traffic-flows-labeled-with-87-apps

Song, R., & Liu, F. (2014, November). Real-time anomaly traffic monitoring based on dynamic k-NN cumulative-distance abnormal detection algorithm. In *2014 IEEE 3rd International Conference on Cloud Computing and Intelligence Systems* (pp. 187-192). IEEE. doi: 10.1109/CCIS.2014.7175727

Song, W., Beshley, M., Przystupa, K., Beshley, H., Kochan, O., Pryslupskyi, A., ... & Su, J. (2020). A software deep packet inspection system for network traffic analysis and anomaly detection. *Sensors*, *20*(6), 1637. doi: 10.1109/DRCN.2015.7149025

Vikram, A. (2020, June). Anomaly detection in network traffic using unsupervised machine learning approach. In 2020 5th International Conference on Communication and Electronics Systems (ICCES) (pp. 476-479). IEEE doi: 10.1109/ICCES48766.2020.9137987

Zou, M., Wang, C., Li, F., & Song, W. (2018). Network phenotyping for network traffic classification and anomaly detection. *arXiv preprint arXiv:1803.01528*. https://doi.org/10.48550/arXiv.1803.01528

Zhao, S., Chandrashekar, M., Lee, Y., & Medhi, D. (2015, March). Real-time network anomaly detection system using machine learning. In *2015 11th international conference on the design of reliable communication networks (drcn)* (pp. 267-270). IEEE. doi: 10.1109/DRCN.2015.7149025