Machine Learning for the Identification of Network Anomalies

Naresh Thoutam, Mayur Sonawane, Ghanshyam Chaudhari, Om Kathe, Prajwal Sontakke
Department of Computer Engineering
Sandip Institute of Technology & Research Centre®, Nashik, India

I. Abstract

The most popular technique for identifying and blocking malicious network requests is the intrusion detection system, or IDS for short. They are positioned carefully to keep an eye on network traffic going to and coming from every device. Most networking devices can employ an IDS with the use of virtual machines and sophisticated switches. While having good accuracy, the classic SIDS (Signature-Based Intrusion Detection System) cannot identify many modern incursions, such as zero-day attacks, as it relies on a pattern matching technique. Instead, the majority of recently launched attacks can be detected using machine learning, statistical, and knowledge-based methods. An anomaly is defined as any significant difference between the observed behavior and the model.The training phase and the testing phase make up the two stages of the development of these models. During the training phase, a model of typical behavior is learned using the average traffic profile. The system's ability to generalize to as-yet-undiscovered intrusions is then determined during the testing step using a fresh data set. In order to identify network traffic anomalies, we have used an unsupervised machine-learning approach called Isolation Forest in this paper. Using the anomaly score,

the algorithm finds the outliers. The KDD data set, a well-known benchmark in the study of Intrusion Detection methods, has been used for training and testing.

Keywords: anomaly detection; machine learning; network security

II. Introduction

Global digitalization is rapidly increasing the number of networking devices, which has grown significantly as well. According to the analysis, the global market for network devices is currently valued at USD 26.4 billion and is predicted to increase at a 6.6% compound annual growth rate (CAGR) through 2027. These networking tools are in charge of transporting various types of public and private data. As a result, the quantity of unidentified attacks has likewise dramatically increased. We require a reliable and effective method to recognise such attacks in order to combat them. There has long been an IDS (intrusion detection system) .They can detect these requests to a certain degree. The foundation of conventional IDS is a signature-based approach to detection. Signature-based IDS can only identify attacks that were previously known about or that have already affected the company because they have a pre-programmed list of known threats and their indications of compromise (IOCs).Due to the extensive use of machine learning-based systems, evaluating it with a regular processor is almost impossible. Machine learning-based systems are used to discover unforeseen dangers to the organization. Anomaly Based Intrusion Detection Systems are the name given to these systems. A trained machine-learning model is given the network traffic data in this case, and it finds the unusual network requests. A network request with unusual parameters is classified as an intrusion and may be harmful. R2L, DoS, U2R, and Probe attack types can all be found by these systems. In anomaly-based intrusion detection systems, supervised and unsupervised machine learning are the two main techniques. Some often used algorithms include ANN, SVM, KNN, and K-Means. The extensive range of options makes a data-driven strategy more precise and effective.

III. Terminologies

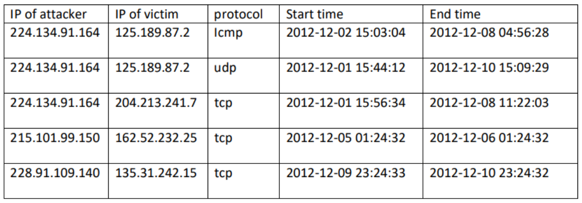
1. **Isolation Forest :**
   IF constructs an ensemble of random trees for a given dataset, with anomalies represented as points in the tree structures. It is helpful when there are a sizable number of unbalanced and dispersed data points. This is due to the fact that outlier data points were easier to discern from typical data points

2. **SVM:**
   Scalable Vector Machines (SVM) are machine learning models useful for finding abnormalities in very unbalanced data sets.

IV. Experimental Analysis

Dataset:

The provided data sets consist of 3 parts: training, validation and test data. Training and test dataset do not contain labels since the purpose of the task is unsupervised anomaly detection. However, for the purpose of tuning hyperparameters, validation dataset was provided with labels. These datasets were first loaded in Splunk for overview and inspection of data and later used in Python for generating features and applying them to models.



IP of attacker	IP of victim	protocol	Start time	End time
224.134.91.164	125.189.87.2	Icmp	2012-12-02 15:03:04	2012-12-08 04:56:28
224.134.91.164	125.189.87.2	udp	2012-12-01 15:44:12	2012-12-10 15:09:29
224.134.91.164	204.213.241.7	tcp	2012-12-01 15:56:34	2012-12-08 11:22:03
215.101.99.150	162.52.232.25	tcp	2012-12-05 01:24:32	2012-12-06 01:24:32
228.91.109.140	135.31.242.15	tcp	2012-12-09 23:24:33	2012-12-10 23:24:32

Summary of the attack

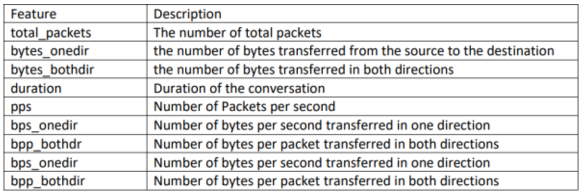
Features:

In the implementation, only one feature (feature 3) is used for prediction. For analysis three features are used with both Isolation Forest and OCSVM, and results are analyzed.
The provided dataset had 14 columns which contained categorical and numeric values. Based on these existing columns, some were modified/processed, removed and added.

In this project, 3 types of different methods for feature generation were used. For all feature sets, PCA and Standardscaler were applied. PCA was used for dimensionality reduction, so the training process gets faster and generalizes models more preventing overfitting. Regarding feature selection, at first only numeric fields were chosen and generating other fields based on them (pps, bps, bpp). And then Feature 2 incorporated one-hot encoded categorical feature
into this. Feature 1 and feature 2 are based on data from each row in the dataset so it generates the same number of lines as the original dataset. However, feature 3 was generated grouped by 5 columns so the new dataset contains fewer rows than the original dataset.

Feature 1: Numeric value (existing + newly generated) + Standardscaler + PCA

The features were overviewed and generated via a query in Splunk first, as seen below.



Feature	Description
total_packets	The number of total packets
bytes_onedir	the number of bytes transferred from the source to the destination
bytes_bothdir	the number of bytes transferred in both directions
duration	Duration of the conversation
pps	Number of Packets per second
bps_onedir	Number of bytes per second transferred in one direction
bpp_bothdr	Number of bytes per packet transferred in both directions
bps_onedir	Number of bytes per second transferred in one direction
bpp_bothdir	Number of bytes per packet transferred in both directions

Feature 2: Feature1 + One-hot encoded categorical feature

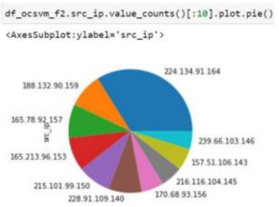
In addition to feature 1, categorical features (src_ip, dst_ip, src_port, dst_port, protocol) were added. Only top 10 occurring encoded features were used since I expected a memory problem if all categorical features were one-hot encoded.

Feature 3: Scale(Cumulative features grouped by stream_id + time-based feature) + PCA

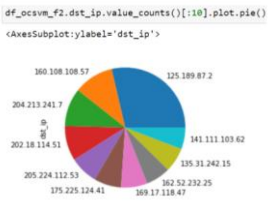
The last feature was generated grouped by (src_ip, dst_ip, src_port, dst_port, protocol) thus, it contains unique rows based on these columns. The idea of grouping was derived from the lecturer's comment on the discussion board. Basically, Feature 1 was summed/averaged and grouped by the 5 columns above. And time- based features (cnt_timed_src, cnt_timed_dst, cnt_timed_srcdst) were added because anomaly traffic may be concentrated in a very short period of time.

Training Dataset Results

1] OCSVM Feature 2

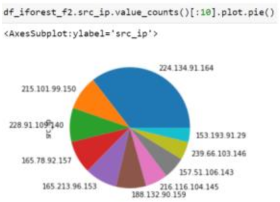


```
df_ocsvm_f2.src_ip.value_counts()[:10].plot.pie()
<AxesSubplot:ylabel='src_ip'>
```

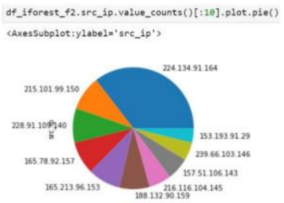


```
df_ocsvm_f2.dst_ip.value_counts()[:10].plot.pie()
<AxesSubplot:ylabel='dst_ip'>
```

2] Isolation Forest Feature 2



```
df_iforest_f2.src_ip.value_counts()[:10].plot.pie()
<AxesSubplot:ylabel='src_ip'>
```



```
df_iforest_f2.src_ip.value_counts()[:10].plot.pie()
<AxesSubplot:ylabel='src_ip'>
```

V. FUTURE SCOPE

Mixing machine learning approaches and developing a hybrid model can improve the model's accuracy even more. Feature normalization can also enhance accuracy. Several feature selection approaches may be used to choose certain qualities that can have a stronger influence on results. Deep learning algorithms have been found to be more robust and accurate. Because of the increase in internal assaults on businesses, it is increasingly critical to assess behavior and discover abnormalities in real-time with great efficiency. Machine learning techniques and analytics of user and entity behavior can be utilized to do this.Both supervised, and unsupervised machine learning may be employed to develop a hybrid system that produces superior results. Parallelization is the typical approach to performance problems in computer science. In the future, the model might be improved to integrate real-time data and identify attacks depending on changes in network traffic.

VI. CONCLUSION

Because of the very uneven data, an unsupervised machine-learning model was created. The AUC computed is 98.3%. The "n estimators" option was set to 100. The parameter value for "contamination" was 0.04, representing 4% of all samples. As the number of different network attacks grows, corporations are developing intrusion detection systems (IDS) that are not only effective but also capable of detecting threats in real time. Anomaly detection is a promising technique in this industry because of its capacity to detect irregularities with low rates of false positive and false negative detection. During implementation, it was revealed that utilizing various values for the available parameters for these algorithms might improve the anomaly identification process. Hence, it might be assumed that a produces superior results.Therefore, it is possible to infer that larger and cleaner data collections provide better results. The contamination parameter is critical in influencing the likelihood of finding

irregularities. Being that greater and cleaner machine data collecting. The contamination parameter is critical in influencing the likelihood of finding irregularities. While machine learning and deep learning applications are still in their early stages in the field of network security, there are still scalability and effectiveness difficulties.

REFERENCES

[1] G. Karatas et al., "Deep Learning in Intrusion Detection Systems" 2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT), Turkey, 2018.

[2] H. Azwar et al., "Intrusion Detection in secure network for Cybersecurity systems using Machine Learning" 2018 IEEE 5th International Conference on Engineering Technologies and Applied Sciences ,Bangkok, Thailand, 2018.

[3] Y. Chang et al., "Network Intrusion Detection Based on Random Forest and Support Vector Machine," IEEE International Conference on Computational Science and Engineering (CSE), Guangzhou, 2017.

[4] Brao, Bobba et al., "Fast kNN Classifiers for Network Intrusion Detection System", Indian Journal of Science and Technology. 2017.

[5] M. Z. Alom et all., "Network intrusion detection for cyber security using unsupervised deep learning approaches", 2017 IEEE National Aerospace and Electronics Conference (NAECON), Dayton, OH, 2017.

[6] Mukkamala et al., "Intrusion detection using neural networks and support vector machines", International Joint Conference 2012.

[7] Azwar, Hassan et all.,"Intrusion Detection in secure network for Cybersecurity systems using Machine Learning and Data Mining", 2018.

[8] Jeya, P et al., "Efficient Classifier for R2L and U2R Attacks", International Journal Comput. Appl. (2012)

[9] Mohana, NK Srinath "Trust Based Routing Algorithms for Mobile Ad-hoc Network", International Journal of Emerging Technologies and Advanced Engineering (IJETAE), volume 2, issue 8, pp. 218-224, IJETAE.

[10] CV Krishna et al. "A Review of Artificial Intelligence Methods for Data Science and Data Analytics: Applications and Research Challenges," International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2018.

[11] F. T. Liu et al., "Isolation Forest," 2008 Eighth IEEE International Conference on Data Mining, Pisa, 2008.

[12] Zhangyu Cheng et al., "Outlier detection using isolation forest and local outlier factor", Conference on Research in Adaptive and Convergent Systems (RACS '19). Association for Computing Machinery, USA.

[13] Yang, Meng et al., "Deep Learning and One-class SVM based Anomalous Crowd Detection", IJCNN.2019.

[14] Ge, Mengmeng et al., "Deep Learning-Based Intrusion Detection for IoT Networks", 2019 IEEE 24th Pacific Rim International Symposium on Dependable Computing (PRDC), pp. 256-25609. IEEE, 2019.

[15] Sathesh, A. (2019). ENHANCED SOFT COMPUTING APPROACHES FOR INTRUSION DETECTION SCHEMES IN SOCIAL MEDIA NETWORKS. Journal of Soft Computing Paradigm (JSCP), 1(02), 69-79.