International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 11 | November - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Detection of Anomalies in Network using Machine Learning

Prof.Naresh Thoutam (Guide)

Author: Mayur Sonawane, Ghanshyam Chaudhari, Om kathe, Prajwal Sontakke

Department of Computer Engineering

Sandip Institute of Technology & Research Center®, Nashik, India

Abstract

The Intrusion Detection System, also known as IDS, is the most widely used method to detect and filter malicious network requests. They are placed at a strategic point to monitor traffic to and from all devices on the network. With the help of VMs and advanced switches, most networking devices can use an IDS. Though the traditional Signature Based Intrusion Detection System known as SIDS gives good accuracy, it cannot detect many new-age intrusions such as zero-day attacks as SIDS is based on pattern matching technique. Instead, machine learning, statistical-based, and knowledge-based methods can detect most newly founded attacks. Any significant deviation between the observed behavior and the model is considered an anomaly. The development of these models comprises two phases: the training phase and the testing phase. In the training phase, the average traffic profile is used to learn a model of normal behavior. Then in the testing phase, a new data set is used to establish the system's capacity to generalize to previously unseen intrusions. In this paper, we have used an unsupervised machine-learning algorithm called Isolation Forest to detect anomalies in network traffic. The algorithm detects the outliers based on the anomaly score. To train and test, the KDD data set has been used, which is a well-known benchmark in the research of Intrusion Detection techniques.

Keywords: industrial control systems; anomaly detection; machine learning; network security

Literature Survey

Fei Tony Liu and Kai Ming Ting, students at Monash University, Victoria,Australia in 2008, first proposed the Isolation Forest Algorithm for outlier detection using special binary trees which they named 'iTree'. Gang Chen, Yuan Li Cai, and Juan Shi students at Jiaotong University proposed an alternative algorithm to Isolation Forest, focusing specifically on outlier detection.
Oussama Ghorbel and Abdulfattah M. Obeid published a comparative study between Centralised and Distributed One-class Outliers Detection Classifier (COODC &DOODC) based on Mahalanobis Kernel used for outlier detection in Wireless SensorNetworks (WSNs).
Hamoud Alshammari focused on the outlier detection method based onNon-Negative Matrix Factorization (NMF).Cong Gao researched Anomaly detection frameworks for outlier and pattern anomaly of time series in wireless sensor networks. In this research, he compared five methods of anomaly detection in WSNs which are Statistical-based, Clustering-based, Classification-based Spectral Decomposition-based, Nearest Neighbour-based, andIsolation-based.Abdelaziz Amara Korba focused on the Anomaly-Based Intrusion DetectionSystem for Ad hoc Networks, applying a statistics-based isolation method that isolates the anomalous nodes.Yuwei Sun, Hideya Ochiai researched on supervised neural network and an
expert-knowledge based labeling method, model training was conducted based on

dataset covering two weeks' network traffic, where the first week's data was employed as the training set and the second week's data was used as the validation set.
Zhilu Wang, Yunfeng Ding published a paper on a distribution network anomaly detection algorithm based on variational auto-encoder is proposed to solve the problem of anomaly detection of distribution terminal data. Yuwei Sun, Hideya Ochiai focused on visualization methods for representing network traffic features using raw data. The raw network traffic data was divided into regulated segments. By employing a supervised neural network and an expert-knowledge based labeling method, model training was conducted based on a dataset covering two weeks' network traffic, where the first week's data was employed as the training set and the The second week's data was used as the validation set.

I. Introduction

With rapid digitalization around the globe, the number of networking devices has also increased by an enormous amount. As per the report, the global network devices market is USD 26.4 billion per the report, which is expected to grow at a Compound Annual Growth Rate (CAGR) of 6.6% by 2027. These networking devices are responsible for carrying all kinds of public and private data. As a result, the number of unknown attacks has also increased rapidly. To counter this, we need a robust and efficient technique to identify such attacks. IDS (Intrusion Detection System) (Azwar 2018, [2])(Karastsetel, n.d., [1]) has been there for a long time. They detect such requests up to a certain extent. The traditional IDS are based on the signature-based method of detection. Signature-based IDS have a pre-programmed list of known threats and their indicators of compromise (IOCs); this is why they can only detect the attacks which are already known or have happened to the organization.

Machine learning based systems are used to detect unknown threats to the organization. These systems are known as Anomaly Based Intrusion Detection Systems. Here the network traffic data is provided to a trained machine-learning model, which detects the anomalous network requests. Network requests having abnormal parameters are labeled as intrusions and possible malicious requests. These systems can detect attack types such as R2L, DoS, U2R, and Probe. Two primary methods can be used in anomaly-based intrusion detection systems: Supervised and Unsupervised Machine Learning. ANN, SVM, KNN, and K-Means are some commonly used algorithms. A data-driven approach is more accurate and efficient because of the vast array of possibilities.

II. FUNDAMENTALS AND BASIC TERMINOLOGIES USED

A. *Isolation Forest :*

For a particular dataset, IF creates an ensemble of random trees, with anomalies expressed as points in the tree structures. When there is a significant Unbalance and scattered data points, it's useful. This is because outlier data points were easier to distinguish than conventional data points. (Liuetal, n.d., [11])

B. *SVM:*

Vector Machines (SVM) are machine learning models that are effective in detecting anomalies in highly imbalanced data sets. (Mukkamal etal, n.d. [6])

III. Experimental Analysis

This implementation attempted to detect anomalous points; there may have been false positives, but our main goal was to detect the

majority of anomalous points or malicious assaults. The dataset is massive and complicated, and testing it with a standard processor is nearly difficult due to the vast number of training occurrences. The big dataset necessitates the use of high-performance computers. The experiment necessitated data pre-processing and cleansing of the data collection. The training and test data are divided in the ratio of 20% test data to 80% training data[12]. This challenge did not make use of the cross-validation set. Outliers and abnormalities were detected using Isolation Forest. Various parameter values were utilized and tested to achieve varying degrees of accuracy.

A. Data set

The NSL-KDD is a security researcher-popular implementation of the KDD detecting anomalies dataset. The data set in this implementation provides 38 distinct sorts of assaults that are aggregated into four fundamental attack classes to give a more visible representation of outcomes. These assaults are classified into four types: DOS, r2I, u2R, and prob attack. Dos attacks vary from other kinds of cyberattacks including that they take down a resource, whilst others penetrate a network or system. (*Azwar, Hassan*[7])

B. Data Preprocessing

There are 41 features in the KDD 99 Dataset. The model requires many character values, which are continuous values of floating-point numbers. Normalization and feature scaling are carried out with the use of a typical scaler. Non-numeric categorical data was used to extract some of the derived features. As a result, it also used the sklearn Label Encoder to encode these variables for training purposes.

III. METHODOLOGY & DIAGRAM

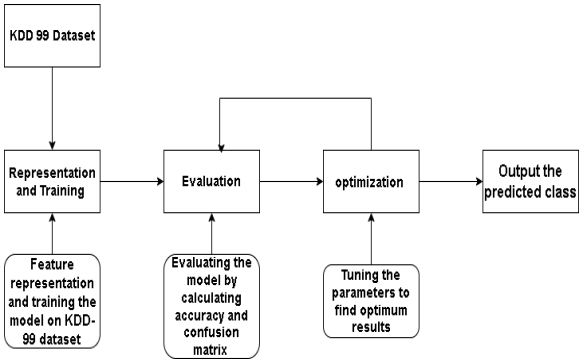

Fig.1. Block diagram of anomaly detection.

The block diagram for anomaly detection is shown in Figure1. Building random forests[5] is how the isolation forest algorithms carry out unsupervised anomaly detection. Afterward, examine the typical depth needed to isolate each point. The Isolation forest model is created and instantiated using a number of parameters. The contamination parameter is the most crucial parameter, which has no bearing on how the model is trained but is essential for interpreting the result. Controls merely whether a scored data point should be regarded as an outlier in the decision function. O(number of samples * n estimators * log sample size) is the time complexity. The isolation forest is particularly effective and well suited for real-time anomaly detection because of its linear complexity.

N estimators: This parameter specifies the number of base estimators and trees that must be constructed for outlier class estimation and detection.

Max_sample: it controls how many training data points are selected to train each tree.

Contamination_param: It represents the percentage of outliers in the provided data set.The contamination factor determines the cutoff point for data points to be deemed abnormal. For our data set, 1% of this was chosen.

Two evaluation metrics were selected: the anomaly score and the AUC score. Since these metrics provide a clear indicator of the model's

effectiveness and robustness, all algorithms in this implementation were assessed based on them. The model's accurate predictions are used to calculate the accuracy score. The model must be taught the presence of anomalous points because it is unsupervised. The workflow for outlier detection is shown in Figure 2.

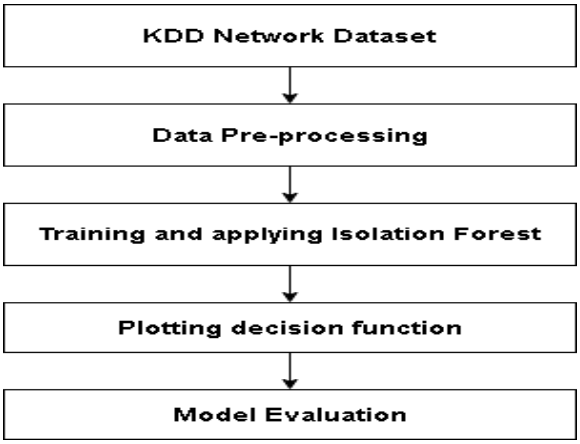

Fig.2. Flowchart of Outlier detection.

C. Training Models

The training and test split was carried out in order to train the algorithms to train independently and test on unidentified samples to obtain better results. To improve performance, a subset of the data set was used. Since the isolation forest is unsupervised, training the model doesn't require target labels. We examined and then provided the various inputs needed to initialize an Isolation forest model.

In comparison to other algorithms, isolation forest is quicker.

$$s\ (x,n) = \ 2\ - E(h(x))$$

IV. RESULTS AND DISCUSSIONS

To increase the precision of our model even further, the results are processed and examined. Dimensionality reduction was accomplished using PCA, which allowed for a clearer depiction of the data set.

A subset of the KDD data set was used in the analysis to train the models. Data from the test split in KDD was randomly chosen for the testing. The models were initially trained on all characteristics, and then the quantity of features chosen was changed to obtain various degrees of accuracy.

The KDD 99 dataset contains information on a variety of assaults, including normal, DoS, probe, and U2R R2. Python and sklearn were used for the experiments.

Any network intrusion detection system's goal is to prevent false positives from occurring as much as possible and to identify all attacks as anomalies. This makes the isolation forest's contamination parameter essential in our use case. The users must be provided with a clear and understandable visualization of these observed anomalies. Depending on the data input, such as the n/w traffic, which varies, the model should peak and exhibit anomalies appropriately.Due to the fact that network traffic differs during weekdays and weekends, it can feature assistive functions that highlight these% differences and display outliers based on

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}}$$

Where c(n) is the average path length of unsuccessful searches in a binary search tree, n is the number of external nodes, and h(x) is the path length for the provided data point after feature splits. An anomaly score is assigned to each observation and subsequent feature value split, and it can be used to make the following decisions: A score of 1 or less denotes anomalies, whereas a score of less than 0.5 is considered benign or normal.

All scores should be close to 0.5 if there are no obvious anomalies across the board.

For high dimensional characteristics like KDD Cup, isolation forests are useful since they are computationally efficient. It is a good option for anomaly detection on the KDD cup data set as a result. Using the data points, the algorithm creates a sizable number of trees, randomly selects certain feature values, then separates those feature values. The underlying assumption is still that an anomalous point has a shorter travel length than inliers. To designate the points as anomalous, the cutoff was selected as a threshold. The crucial finding is that the system accurately identified Normal and Anomaly sites for DOS. The accuracy of less well-known attacks is, however, lower. It has been plotted that some normal points are anomalous.

V. FUTURE SCOPE

Combining the machine learning methods and creating a hybrid model can further increase the model's accuracy. Accuracy can also be improved by feature normalization. To choose specific features that can have a greater impact on results, various feature selection methods can be utilized. Deep learning approaches have been shown to be robust and to have improved accuracy. It is now crucial to evaluate behavior and spot anomalies in real time with high efficiency due to the rise in internal attacks on organizations. Analytics of user and entity behavior and machine learning techniques can be used to accomplish this. To create a hybrid system that can produce better outcomes, supervised and unsupervised machine learning can both be used.The traditional solution to performance issues in computer science is parallelization. The model may be enhanced in the future to incorporate real-time data and suggest assaults based on changes in network traffic.(*Sathesh, A. (2019)* [15])

VI. CONCLUSION

Due to the very unbalanced data, an unsupervised machine-learning model was developed. The calculated AUC score is 98.3%. We left the "n estimators" argument at 100. The "contamination" parameter value was 0.04, representing 4% of all samples. The number of diverse network attacks is increasing dramatically, thus businesses are creating intrusion detection systems (IDS) that are not only very effective but also able to identify threats in real-time.The ability to detect abnormalities with low rates of false positive and false negative detection makes anomaly detection a promising technology in this field. During implementation, it was discovered that using different values of the available parameters for these algorithms could enhance the anomaly detection process. Additionally, it may be inferred that better outcomes are produced by a data collection that is larger and cleaner. When determining the probability of detecting abnormalities, the contamination parameter is crucial. Being that machine learning and deep learning applications are still relatively new to the field of network security, there are still issues with scalability and effectiveness.(*CV Krishna* [10])
(*Ge, Mengmeng* [14])

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.