Journal of Engineering Sciences

Vol 13, Issue 06, JUNE /2022

ISSN NO:0377-9254

The prediction of data with network attack with multiple algorithms

[1]Dr.K.Sreenivasulu

[2]Patnamu Subba Tejaswini

[3]Narahari Supriya

[1]GUIDE [2,3] U.G.SCHOLARS

G PULLAIAH COLLEGE OF ENGINEERING AND TECHNOLOGY

ABSTRACT

In the current network security situation, the types of network threats are complex and changeable. With the development of the Internet and the application of information technology, the general trend is opener. Important data and important business applications will face more serious security threats. However, with the development of cloud computing technology, the trend of large-scale deployment of important business applications in cloud centers has greatly increased. The development and use of software-defined networks in cloud data centers have greatly reduced the effect of traditional network security boundary protection. How to find an effective way to protect important applications in open multi-step large-scale cloud data centers is a problem we need to solve. Threat intelligence has become an important means to solve complex network attacks, realize real-time threat early warning and attack tracking because of its ability to analyze the threat intelligence data of various network attacks. Based on the research of threat intelligence, machine learning, cloud central network, SDN and other technologies, this paper proposes an active defense method of network security based on threat intelligence for super-large cloud data centers.

I. 1.INTRODUCTION

Objective:

This paper mainly includes Threat Intelligence data center and network attack blocking component. As the control center of threat intelligence collection, analysis and strategy distribution, Threat Intelligence data center is deployed in cloud data center. The network attack blocking components are deployed in the form of virtualization components at the Internet entrances and exits of each cloud data center, and the SDN technology is used for traffic processing and policy distribution. The network attack blocking component is deployed at the Internet entrance and exit to monitor, identify and automatically block the malicious attack source, so as to achieve the effect of one point monitoring and whole network blocking.

Problem Statement:

The development and use of software-defined networks in cloud data centers have greatly reduced the effect of traditional

Journal of Engineering Sciences

network security boundary protection. How to find an effective way to protect important applications in open multi-step large-scale cloud data centers is a problem we need to solve.

Project Scope:

This paper studies and designs a cloud platform protection method using Threat Intelligence for active defense. Through the research of key technologies such as big data intelligent analysis, multi-source Threat Intelligence Fusion and SDN technology, we can build an accurate network security early warning and interception capability at the Internet boundary of the cloud center, and realize the linkage protection capability of the whole distributed cloud center by using effective threat intelligence.

This paper proposes a technical scheme of network attack blocking based on Threat Intelligence:

• Constructing the "effectiveness, accuracy" threat intelligence model to adapt to the use of attack interruption.

• Building a new model of border linkage protection adapted to cloud computing architecture.

• Building Threat Intelligence big data center with unified strategy. Through the experimental test, our scheme has higher efficiency and accuracy, and lower false alarm rate.

II. .LITERATURE SURVEY

2.1Cyber Threat Intelligence Model: An Evaluation of Taxonomies, Sharing Standards, and Ontologies within Cyber Threat Intelligence

Abstract: Threat intelligence is the provision of evidence-based knowledge about existing or potential threats. Benefits of threat intelligence include improved efficiency and effectiveness in security operations in terms of detective and preventive capabilities. Successful threat intelligence within the cyber domain demands a knowledge base of threat information and an expressive way to represent this knowledge. This purpose is served by the use of taxonomies, sharing standards, and ontologies. This paper introduces the Cyber Threat Intelligence (CTI) model, which enables cyber defenders to explore their threat intelligence capabilities and understand their position against the ever-changing cyber threat landscape. In addition, we use our model to analyze and evaluate several existing taxonomies, sharing standards, and ontologies relevant to cyber threat intelligence. Our results show that the cyber security community lacks an ontology covering the complete spectrum of threat intelligence. To conclude, we argue the importance of developing a multi-layered cyber threat intelligence ontology based on the CTI model and the steps should be taken under consideration, which are the foundation of our future work.

2.2 Analyzing Malicious URLs using a Threat Intelligence System.

Journal of Engineering Sciences

Abstract: Threat intelligence and management systems form a vital component of an organization's cybersecurity infrastructure. Threat intelligence, when used with active monitoring of network traffic, can be critical to ensure reliable data communication between endpoints. Threat intelligence systems are well suited for analyzing anomalous behaviors in network traffic and can be employed to assist organizations in identifying and successfully responding to cyber-attacks. In this paper, we present a machine learning approach for clustering malicious uniform resource locators (URLs). We focus on a URL dataset gathered from a threat intelligence feeds framework. We implement a k-means clustering solution for grouping malicious URLs obtained from open source threat intelligence feeds. We demonstrate the effectiveness of our unsupervised learning technique to discover the hidden structures in the malicious URL dataset. Our URL keyword/text clustering solution provides valuable insights about the malicious URLs and aids network operators in policy decisions to mitigate cyber-attacks. The clusters obtained using our approach has a silhouette coefficient of 0.383 for a dataset containing over 11,000 malicious URLs. Lastly, we develop a probabilistic scoring model to calculate the percentage of malicious keywords present in a given URL. After analyzing over 72,000 malicious keywords, our model successfully identifies over 80% of the URLs in a test dataset as malicious.

2.3 Preventing Poisoning Attacks On AI Based Threat Intelligence Systems

Abstract: As AI systems become more ubiquitous, securing them becomes an emerging challenge. Over the years, with the surge in online social media use and the data available for analysis, AI systems have been built to extract, represent and use this information. The credibility of this information extracted from open sources, however, can often be questionable. Malicious or incorrect information can cause a loss of money, reputation, and resources; and in certain situations, pose a threat to human life. In this paper, we use an ensembled semi-supervised approach to determine the credibility of Reddit posts by estimating their reputation score to ensure the validity of information ingested by AI systems. We demonstrate our approach in the cybersecurity domain, where security analysts utilize these systems to determine possible threats by analyzing the data scattered on social media websites, forums, blogs, etc.

2.4 Data-driven analytics for cyber-threat intelligence and information sharing

Abstract: Efficient analysis of shared Cyber Threat Intelligence (CTI) information is crucial for network risk assessment and security hardening. There is a growing interest in implementing a proactive line of defense through threat profiling. However,

Journal of Engineering Sciences

determining the resiliency of a particular network with respect to relevant threats reported in CTI shared data remains a challenge, largely due to the lack of semantics and contextual information present in textual representations of the threat knowledge. To overcome the limitations of existing CTI frameworks, we devise a threat analytics framework based on Web Ontology Language (OWL) for formal specification, semantic reasoning, and contextual analysis, allowing the derivation of network associated threats from large volumes of shared threat feeds. Our ontology represents constructs of Structured Threat Information eXpression (STIX) with the additional concepts of Cyber Observable eXpression (CybOX), network configurations, and Common Vulnerabilities and Exposure (CVE) for risk analysis and threat actor profiling. The framework provides an automated mechanism to investigate cyber threats targeting the network under question by classifying the threat relevance, determining threat likelihood, identifying the affected and exposed assets through formulated rules and inferences. We perform a comprehensive structural and conceptual evaluation of critical advanced persistent threats (APTs) collected from credible sources and determine their relevance and risk posed to realistic network case studies. Finally we show that the proposed framework is novel in the type of analytics it provides and outperforms other competing approaches in terms of efficiency and effectiveness.

2.5 Towards automated threat intelligence fusion

AUTHORS: A. Modi, Z. Sun, A. Panwar, T. Khairnar, Z. Zhao, A. Doup, G. Ahn, and P. Black

ABSTRACT: The volume and frequency of new cyber attacks have exploded in recent years. Such events have very complicated workflows and involve multiple criminal actors and organizations. However, current practices for threat analysis and intelligence discovery are still performed piecemeal in an ad-hoc manner. For example, a modern malware analysis system can dissect a piece of malicious code by itself. But, it cannot automatically identify the criminals who developed it or relate other cyber attack events with it. Consequently, it is imperative to automatically assemble the jigsaw puzzles of cybercrime events by performing threat intelligence fusion on data collected from heterogeneous sources, such as malware, underground social networks, cryptocurrency transaction records, etc. In this paper, we propose an Automated Threat Intelligence fuSion framework (ATIS) that is able to take all sorts of threat sources into account and discover new intelligence by connecting the dots of apparently isolated cyber events. To this end, ATIS consists of 5 planes, namely analysis, collection, controller, data and application planes. We discuss the design choices we made in the function of each plane and the interfaces between

Journal of Engineering Sciences

two adjacent planes. In addition, we develop two applications on top of ATIS to demonstrate its effectiveness.

III. .SYSTEM ANALYSIS

3.1 EXISTING SYSTEM:

The development and deployment of software-defined networks in cloud data centres have drastically weakened the influence of traditional network security boundary protection. How to discover an effective solution to safeguard vital applications in open multi-step large-scale cloud data centres is a challenge we need to tackle.

3.1.1 DISADVANTAGES OF EXISTING SYSTEM:

- Network and production network through the Internet. Due to the trend of open interconnection, the network attack plane has been greatly expanded,

3.2 PROPOSED SYSTEM:

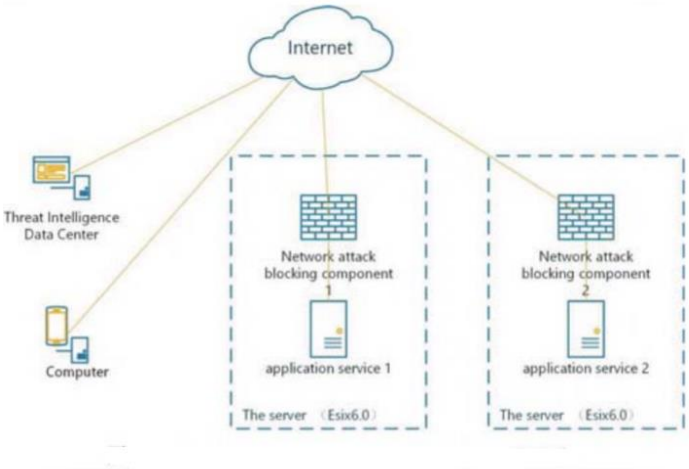
- In this project,the trend of large-scale deployment of important business applications in cloud centers has greatly increased. The development and use of software-defined networks in cloud data centers have greatly reduced the effect of traditional network security boundary protection. How to find an effective way to protect important applications in open multi-step large-scale cloud data centers is a problem we need to solve. Threat intelligence has become an important means to solve complex network attacks, realize real-time threat early warning and attack tracking because of its ability to analyze the threat intelligence data of various network attacks. Based on the research of threat intelligence, machine learning, cloud central network, SDN and other technologies, this paper proposes an active defense method of network security based on threat intelligence for super-large cloud data centers.

3.2.1 ADVANTAGES OF PROPOSED SYSTEM:

- it proposes an active defense method of network security based on threat intelligence for super-large cloud data centers.

IV. SYSTEM DESIGN

4.1 SYSTEM ARCHITECTURE:



Journal of Engineering Sciences

V. MODULES:

- upload Train Dataset
- Run Preprocessing TF-IDF Algorithm
- Generate Event Vector
- Neural Network Profiling
- Run SVM Algorithm
- Run KNN Algorithm
- Run Naive Bayes Algorithm
- Run Decision Tree Algorithm
- Accuracy Comparison Graph
- Precision Comparison Graph
- Recall Comparison Graph
- FMeasure Comparison Graph

MODULES DESCRIPTION:

Propose algorithms consists of following module

1) Data Parsing: This module take input dataset and parse that dataset to create a raw data event model
2) TF-IDF: using this module we will convert raw data into event vector which will contains normal and attack signatures
3) Event Profiling Stage: Processed data will be splitted into train and test model based on profiling events.
4) Deep Learning Neural Network Model: This module runs CNN and LSTM algorithms on train and test data and then generate a training model. Generated trained model will be applied on test data to calculate prediction score, Recall, Precision and FMeasure. Algorithm will learn perfectly will yield better accuracy result and that model will be selected to deploy on real system for attack detection.

Datasets which we are using for testing are of huge size and while building model it's going to out of memory error but kdd_train.csv dataset working perfectly but to run all algorithms it will take 5 to 10 minutes. You can test remaining datasets also by reducing its size or running it on high configuration system.

VI. ALGORITHM

k nearest neighbor algorithm

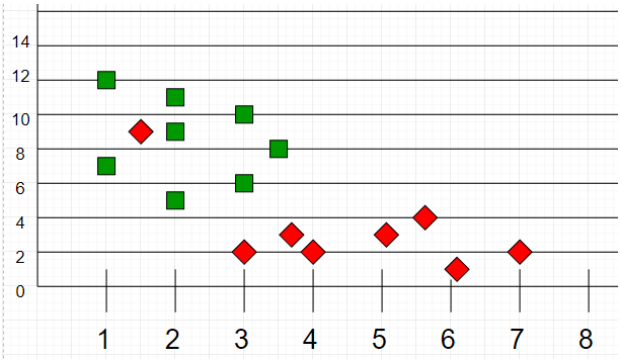
K-Nearest Neighbors is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection.

It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as GMM, which assume a Gaussian distribution of the given data).

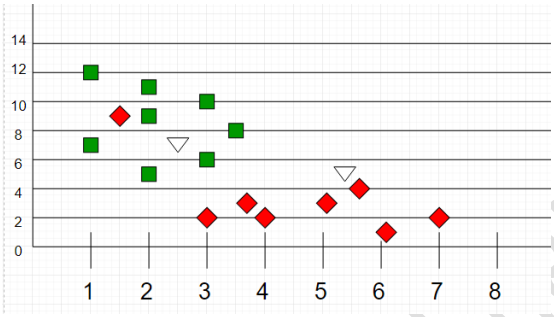
We are given some prior data (also called training data), which classifies coordinates into groups identified by an attribute.

As an example, consider the following table of data points containing two features:

Journal of Engineering Sciences



Now, given another set of data points (also called testing data), allocate these points a group by analyzing the training set. Note that the unclassified points are marked as 'White'.



Naive Bayes Classifier: Naive Bayes is a classification technique with a notion which defines all features are independent and unrelatedto each other. It defines that status of a specific feature in a class does not affect the status of another feature. Sinceit is based on conditional probability it is considered as a powerful algorithm employed for classification purpose.It works well for the data with imbalancing problems and missing values. Naive Bayes is a machine learningclassifier which employs the Bayes Theorem. Using Bayes theorem posterior probability

P(C|X) can be calculated from P(C),P(X) and P(X|C).

Therefore,P(C|X)=(P(X|C)

P(C))/P(X)

Where,P(C|X)=target class's posterior probability .

P(X|C)=predictor class's probability.

P(C)=class C's probability being true.

P(X)=predictor's prior probability.

Decision Tree Classifier:

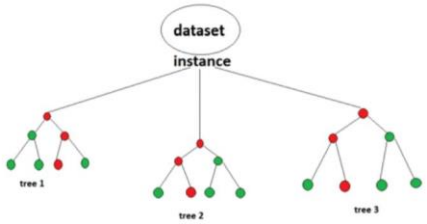
Decision Tree is a supervised machine learning algorithm used to solve classification problems. The main objectiveof using Decision Tree in this research work is the prediction of target class using decision rule taken from prior data.It uses nodes and internodes for the prediction and classification. Root nodes classify the instances with differentfeatures. Root nodes can have two or more branches while the leaf nodes represent classification. In every stage, Decision tree chooses each node by evaluating the highest information gain among all the attributes. The evaluated performance of Decision Tree technique

Random Forest Algorithm

Random Forest algorithm is a supervised classification algorithm. We can see it from its name, which is to create a forest by some way and make it random. There is a direct relationship between the number of trees in the forest and the results it can get: the larger the number of trees, the more accurate the result. But one thing to note is that creating

Journal of Engineering Sciences

the forest is not the same as constructing the decision with information gain or gain index approach. The decision tree is a decision support tool. It uses a tree-like graph to show the possible consequences. If you input a training dataset with targets and features into the decision tree, it will formulate some set of rules. These rules can be used to perform predictions. When we have our dataset categorized into 3 category so now Random forest helps to make classes from the dataset. Random forest is clusters of decision trees all together, if you input a training dataset with features and labels into a decision tree, it will formulate some set of rules, which will be used to make the predictions.



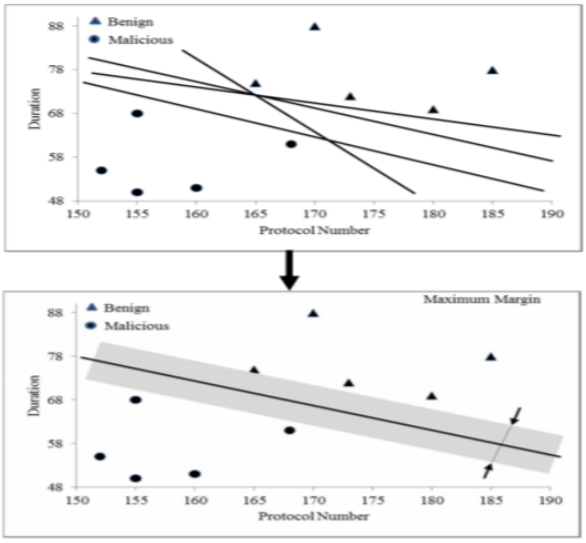
SUPPORT VECTOR MACHINE(SVM)

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well (look at the below snapshot). The SVM algorithm is implemented in practice using a kernel. The learning of the hyperplane in linear SVM is done by transforming the problem using some linear algebra, which is out of the scope of this introduction to SVM. A powerful insight is that the linear SVM can be rephrased using the inner product of any two given observations, rather than the observations themselves. The inner product between two vectors is the sum of the multiplication of each pair of input values. For example, the inner product of the vectors [2, 3] and [5, 6] is 2*5 + 3*6 or 28. The equation for making a prediction for a new input using the dot product between the input (x) and each support vector (xi) is calculated as follows:

$$f(x) = B0 + sum(ai * (x,xi))$$

This is an equation that involves calculating the inner products of a new input vector (x) with all support vectors in training data. The coefficients B0 and ai (for each input) must be estimated from the training data by the learning algorithm.

Journal of Engineering Sciences

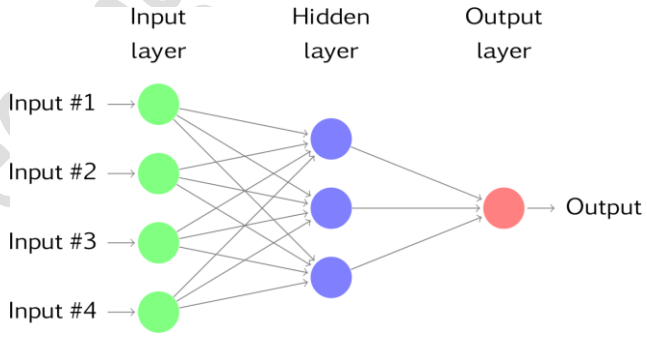


CNN:

To demonstrate how to build a convolutional neural network based image classifier, we shall build a 6 layer neural network that will identify and separate one image from other. This network that we shall build is a very small network that we can run on a CPU as well. Traditional neural networks that are very good at doing image classification have many more parameters and take a lot of time if trained on normal CPU. However, our objective is to show how to build a real-world convolutional neural network using TENSORFLOW.

Neural Networks are essentially mathematical models to solve an optimization problem. They are made of neurons, the basic computation unit of neural networks. A neuron takes an input (say x), do some computation on it (say: multiply it with a variable w and adds another variable b) to produce a value (say; z= wx+b). This value is passed to a non-linear function called activation function (f) to produce the final output(activation) of a neuron. There are many kinds of activation functions. One of the popular activation function is Sigmoid. The neuron which uses sigmoid function as an activation function will be called sigmoid neuron. Depending on the activation functions, neurons are named and there are many kinds of them like RELU, TanH.

If you stack neurons in a single line, it's called a layer; which is the next building block of neural networks. See below image with layers



To predict image class multiple layers operate on each other to get best match layer and this process continues till no more improvement left.

LSTM:

Long short-term memory (**LSTM**) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. It can not only process single data points (such as images), but also entire sequences of data (such as speech or video). For example, LSTM is applicable to tasks such

as unsegmented, connected handwriting recognition, speech recognition[3][4] and anomaly detection in network traffic or IDSs (intrusion detection systems).

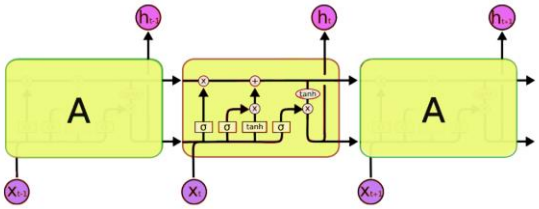
A common LSTM unit is composed of a **cell**, an **input gate**, an **output gate** and a **forget gate**. The cell remembers values over arbitrary time intervals and the three *gates* regulate the flow of information into and out of the cell.

LSTM networks are well-suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series. LSTMs were developed to deal with the vanishing gradient problem that can be encountered when training traditional RNNs. Relative insensitivity to gap length is an advantage of LSTM over RNNs, hidden Markov models and other sequence learning methods in numerous applications

Training:

An RNN using LSTM units can be trained in a supervised fashion, on a set of training sequences, using an optimization algorithm, like gradient descent, combined with backpropagation through time to compute the gradients needed during the optimization process, in order to change each weight of the LSTM network in proportion to the derivative of the error (at the output layer of the LSTM network) with respect to corresponding weight.

A problem with using gradient descent for standard RNNs is that error gradients vanish exponentially quickly with the size of the time lag between important events. However, with LSTM units, when error values are back-propagated from the output layer, the error remains in the LSTM unit's cell. This "error carousel" continuously feeds error back to each of the LSTM unit's gates, until they learn to cut off the value.



VII CONCLUSION

Through the research of threat intelligence technology processing technology, this paper puts forward a method of network border active defense based on Threat Intelligence, which can effectively reduce the security protection pressure of business application, and quickly improve the overall network security protection level. In the next step, it is necessary to conduct in-depth research on the accuracy and adaptability of threat information, build a scientific and reasonable deep learning model, and reduce the risk of network security Small false blocking rate of border, building an efficient border protection capability of joint defense and joint control.

VIII. REFERENCES

[1] Mavroeidis, V., & Bromander, S.. (2017). Cyber Threat Intelligence Model: An Evaluation of Taxonomies, Sharing Standards, and Ontologies within Cyber Threat Intelligence. 2017 European

Journal of Engineering Sciences

Intelligence and Security Informatics Conference (EISIC). IEEE.

[2] Sezer S, Scott-Hayward S, Chouhan P, et al. Are we ready for SDN? Implementation challenges for software-defined networks[J]. IEEE Communications Magazine, 2013, 51(7):36-43.

[3] Nayak, S., Nadig, D., & Ramamurthy, B.. (2019). Analyzing Malicious URLs using a Threat Intelligence System. 2019 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS). IEEE.

[4] Khurana N, Mittal S, Piplai A, et al. Preventing Poisoning Attacks On AI Based Threat Intelligence Systems[C]// 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, 2019.

[5] Singh A, Gukal S. Using high-interaction networks for targeted threat intelligence^]. 2019.

[6] Qamar, S., Anwar, Z., Rahman, M. A., Al-Shaer, E., & Chu, B. T.. (2017). Data-driven analytics for cyber-threat intelligence and information sharing. Computers & Security, 67(JUN.), 35-58.

[7] Tax D M J, Duin R P W. Support vector domain description[J]. Pattern recognition letters, 1999, 20(11-13): 1191-1199.

[8] Xiaoyan R, Danwa S. Research on Cyber-Attack Defense System Based on Big Data and Threat Intelligence^]. Journal of Information Security Research, 2019. [9] S. Lee and T. Shon, Open source intelligence base cyber threat inspection framework for critical infrastructures, in 2016 Future Technologies Conference (FTC), Dec 2016, pp. 1030-1033.

Journal of Engineering Sciences