A User Centric Machine Learning Framework For Cyber Security Operation Center

1. Prof.R.Yadagiri Rao,professor,Head,H&S,Sri Indu Institute of Engineering&Technology(SIIET), Sheriguda,Ibrahimpatnam,Hydarabad,

2.M.Sruthi , Asso professor,CSE,SIIET,Sheriguda,Ibrahimpatnam,Hydarabad,

msruthi1248@gmail.com

3.Pravalika,Student,CSE,SIIET,Sheriguda,Ibrahimpatnam,Hydarabad

4.Deepika,Student,CSE,SIIET,Sheriguda,Ibrahimpatnam,Hydarabad

5.Sumedha,Student,CSE,SIIET,Sheriguda,Ibrahimpatnam,Hydarabad

6.Mounika,Student,CSE,SIIET,Sheriguda,Ibrahimpatnam,Hydarabad

1. ABSTRACT:

In order to ensure a company's Internet security, SIEM (Security Information and Event Management) system is in place to simplify the various preventive technologies and flag alerts for security events. Inspectors (SOC) investigate warnings to determine if this is true or not. However, the number of warnings in general is wrong with the majority and is more than the ability of SCO to handle all awareness. Because of this, malicious possibility. Attacks and compromised hosts may be wrong. Machine learning is a possible approach to improving the wrong positive rate and improving the productivity of SOC analysts. In this article, we create a user-centric engineer learning framework for the Internet Safety Functional Center in the real organizational context. We discuss regular data sources in SOC, their work flow, and how to process this data and create an effective machine learning system. This article is aimed at two groups of readers. The first group is intelligent researchers who have no knowledge of data scientists or computer safety fields but who engineer should develop machine learning systems for machine safety. The second groups of visitors are Internet security practitioners that have deep knowledge and expertise in Cyber Security, but do Machine learning experiences do not exist and I'd like to create one by themselves. At the end of the paper, we use the account as an example to demonstrate full steps from data collection, label creation, feature engineering, machine learning algorithm and sample performance evaluations using the computer built in the SOC production of Seyondike.

Key words: GBM, boosting algorithm, Heart performance.

Copyright @ 2021 Authors

I INTRODUCTION

Cyber security incidents will cause significant financial and reputation impacts on enterprise. In order to detect malicious activities, the SIEM (Security Information and Event Management) system is built in companies or government. The system correlates event logs from endpoint, firewalls, IDS/IPS (Intrusion Detection/Prevention System), DLP (Data Loss Protection), DNS (Domain Name System), DHCP (Dynamic Host Configuration Protocol), Windows/Unix security events, VPN logs etc. The security events can be grouped into different categories [1]. The logs have terabytes of data each day. From the security event logs, SOC (Security Operation Center) team develops so-called use cases with a pre-determined severity based on the analysts¶ experiences. They are typically rule based correlating one or more indicators from different logs. These rules can be network/host based or time/frequency based. If any pre-defined use case is triggered, SIEM system will generate an alert in real time. SOC analysts will then investigate the alerts to decide whether the user related to the alert is risky (a true positive) or not (false positive). If they find the alerts to be suspicious from the analysis, SOC analysts will create OTRS (Open Source Ticket Request System) tickets. After initial investigation, certain OTRS tickets will be escalated to tier 2 investigation system (e.g., Co3 System) as severe security incidents for further investigation and remediation by Incident Response Team. However, SIEM typically generates a lot of the alerts, but with a very high false positive rate. The number of alerts per day can be hundreds of thousands, much more than the capacity for the SOC to investigate all of them. Because of this, SOC may choose to investigate only the alerts with high severity or suppress the same type of alerts. This could potentially miss some severe attacks. Consequently, a more intelligent and automatic system is required to identify risky users. The machine learning system sits in the middle of SOC work flow, incorporates different event logs, SIEM alerts and SOC analysis results and generates comprehensive user risk score for security operation center. Instead of directly digging into large amount of SIEM alerts and trying to find needle in a haystack, SOC analysts can use the risk scores from machine learning system to prioritize their investigations, starting from the users with highest risks. This will greatly improve their efficiency, optimize their job queue management, and ultimately enhance Specifically, our approach constructs a framework of

Copyright @ 2021 Authors

usercentric machine learning system to evaluate user risk based on alert information. This approach can provide security analyst a comprehensive risk score of a user and security analyst can focus on those users with high risk scores. To the best of our knowledge, there is no previous research on building a complete systematic solution for this application. The main contribution of this paper is as follows: x An advanced user-centric machine learning system is proposed and evaluated by real industry data to evaluate user risks. The system can effectively reduce the resources to analyze alerts manually while at the same time enhance enterprise security. x A novel data engineering process is offered which integrates alert information, security logs, and SOC analysts¶ investigation notes to generate features and propagate labels for machine learning models.

II EXISTING SYSTEM

Most approaches to security in the enterprise have focused on protecting the network infrastructure with no or little attention to end users. As a result, traditional security functions and associated devices, such as firewalls and intrusion detection and prevention devices, deal mainly with network level protection. Although still part of the overall security story, such an approach has limitations in light of the new security challenges described in the previous section.

Data Analysis for Network Cyber-Security focuses on monitoring and analyzing network traffic data, with the intention of preventing, or quickly identifying, malicious activity. Risk values were introduced in an information security management system (ISMS) and quantitative evaluation was conducted for detailed risk assessment. The quantitative evaluation showed that the proposed countermeasures could reduce risk to some extent. Investigation into the cost-effectiveness of the proposed countermeasures is an important future work.It provides users with attack information such as the type of attack, frequency, and target host ID and source host ID. Ten et al. proposed a cyber-security framework of the SCADA system as a critical infrastructure using real-time monitoring, anomaly detection,

Copyright @ 2021 Authors

and impact analysis with an attack tree-based methodology, and mitigation strategies

DISADVANTAGE:

1. Firewalls can be difficult to configure correctly.
2. Incorrectly configured firewalls may block users from performing actions on the Internet, until the firewall configured correctly.
3. Makes the system slower than before.
4. Need to keep updating the new software in order to keep security up to date.
5. Could be costly for average user.
6. The user is the only constant

III PROPOSED SYSTEM

User-centric cyber security helps enterprises reduce the risk associated with fast-evolving end-user realities by reinforcing security closer to end users. User-centric cyber security is not the same as user security. User-centric cyber security is about answering peoples' needs in ways that preserve the integrity of the enterprise network and its assets. User security can almost seem like a matter of protecting the network from the user — securing it against vulnerabilities that user needs introduce. User-centric security has the greater value for enterprises. cyber-security systems are real-time and robust independent systems with high performances requirements. They are used in many application domains, including critical infrastructures, such as the national power grid, transportation, medical, and defense. These applications require the attainment of stability, performance, reliability, efficiency, and robustness, which require tight integration of computing, communication, and control technological systems. Critical infrastructures have always been the target of criminals and are affected by security threats because of their complexity and cyber-security connectivity. These CPSs face security breaches when people, processes,

Copyright @ 2021 Authors

technology, or other components are being attacked or risk management systems are missing, inadequate, or fail in any way. The attackers target confidential data. Main scope of this project in reduce the unwanted data for the dataset.

ADVANTAGES:

1) Protects system against viruses, worms, spyware and other

2) Protection against data from theft.

3) Protects the computer from being hacked.

4) Minimizes computer freezing and crashes.

5) Gives privacy to users

6) Securing the user-aware network edge

7) Securing mobile users' communications '

8) Managing user-centric security

IV METHODOLOGY:

CYBER ANALYSIS

Cyber threatanalysis is a process in which the knowledge of internal and external information vulnerabilities pertinent to a particular organization is matched against real-world cyber-attacks. With respect to cyber security, this threat-oriented approach to combating cyber-attacks represents a smooth transition from a state of reactive security to a state of proactive one. Moreover, the desired result of a threat assessment is to give best practices on how to maximize the protective instruments with respect to availability, confidentiality and integrity, without turning back to usability and functionality conditions. CYPER ANALYSIS.A threat could be anything that leads to interruption, meddling or destruction of any valuable service or item existing in the firm's repertoire. Whether of "human" or "nonhuman" origin, the analysis must scrutinize each element that may bring about conceivable security risk.

DATASET MODIFICATION

If a dataset in your dashboard contains many dataset objects, you can hide specific dataset objects from display in the Datasets panel. For example, if you decide to

Copyright @ 2021 Authors

import a large amount of data from a file, but do not remove every unwanted data column before importing the data into Web, you can hide the unwanted attributes and metrics, To hide dataset objects in the Datasets panel, To show hidden objects in the Datasets panel, To rename a dataset object, To create a metric based on an attribute, To create an attribute based on a metric, To define the geo role for an attribute, To create an attribute with additional time information, To replace a dataset object in the dashboard

DATA REDUCTION

Improve storage efficiency through data reduction techniques and capacity optimization using datareduplication, compression, snapshots and thin provisioning. Data reduction via simply deleting unwanted or unneeded data is the most effective way to reduce a storing's data

RISKY USER DETECTION

False alarm immunity to prevent customer embarrassment, High detection rate to protect all kinds of goods from theft, Wide-exit coverage offers greater flexibility for entrance/exit layouts, Wide range of attractive designs complement any store décor, Sophisticated digital controller technology for optimum system performance

ALGORITHM:

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

Copyright @ 2021 Authors

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.

ARCHITECTURE

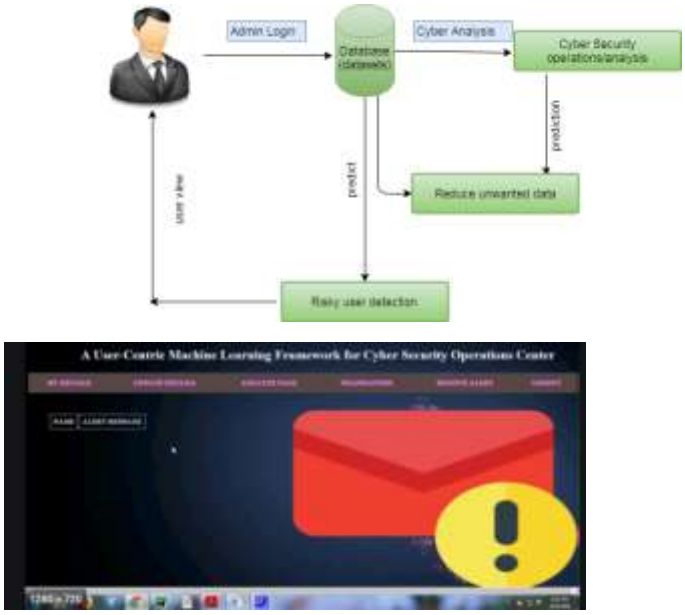

Fig.4.1. Home page.

V CONCLUSION

We provide a user-centered computer learning system that affects large data from various security logs, awareness information, and inspector intelligence. This method provides complete configuration and solution for dangerous user detection for the Enterprise System Operating Center. Select machine learning methods in the SOC product environment, evaluate efficiency, IO, host and users to create user-centric features. . Even with simple mechanical

Copyright @ 2021 Authors

learning algorithms, we prove that the learning system can understand more insights from the rankings with the most unbalanced and limited labels. More than 20% of the neurological model of modeling is 5 times that of the current rule-based system. To improve the detection precision situation, we will examine other learning methods to improve the data acquisition, daily model renewal, real time estimate, fully enhance and organizational risk detection and management. As for future work, let's examine other learning methods to improve detection accuracy.

REFERENCES

[1] SANS Technology Institute. "The 6 Categories of Critical Log Information."2013.

[2] X.Li and B.Lui "Learning to classify text using positive and unlabeled data", Proceedings of the 18th international joint conference on Artificial intelligence, 2003

[3] A. L. Buczak and E. Guven. "A survey of data mining and machine learning methods for cyber security intrusion detection", IEEE Communications Surveys & Tutorials 18.2 (2015): 1153-1176.

[4] S. Choudhury and A. Bhowal. "Comparative analysis of machine learning algorithms along with classifiers for network intrusion detection", Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), 2015.

[5] N. Chand et al. "A comparative analysis of SVM and its stacking with other classification algorithm for intrusion detection", Advances in Computing, Communication, & Automation (ICACCA), 2016.

[6] K. Goeschel. "Reducing false positives in intrusion detection systems using data-mining techniques utilizing support vector machines, decision trees, and naive Bayes for off-line analysis", SoutheastCon, 2016.

[7] M. J. Kang and J. W. Kang. "A novel intrusion detection method using deep neural network for in-vehicle network security", Vehicular Technology Conference, 2016.

Copyright @ 2021 Authors