a
Journal of Engineering Sciences

Vol 13, Issue 06, June/2022

ISSN NO:0377-9254

MACHINE LEARNING TECHNIQUES APPLIED TO DETECT CYBER ATTACKS ON WEB APPLICATIONS

MR.S. Chandra Sekhar , Assistant professor, MTech, Department of CSE

Muvva Sri Nithya, BTech,Department of CSE, 18311A05C0@sreenidhi.edu.in

Arepalli Harshitha,BTech, Department of CSE, 189P1A0561@sreenidhi.edu.in

Buttamgari Ramya Reddy,BTech,Department Of CSE,189P1A0562@sreenidhi.edu.in

Sreenidhi Institute of Science and Technology

ABSTRACT: The increased usage of cloud services, growing number of web applications users, changes in network infrastructure that connects devices running mobile operating systems and constantly evolving network technology cause novel challenges for cyber security. As a result, to counter arising threats, network security mechanisms, sensors and protection schemes also have to evolve, to address the needs and problems of the users. In this article, we focus on countering emerging application layer cyber attacks since those are listed as top threats and the main challenge for network and cyber security. The major contribution of the article is the proposition of machine learning approach to model normal behaviour of application and to detect cyber attacks. The model consists of patterns (in form of Perl Compatible Regular Expressions (PCRE) regular expressions) that are obtained using graph-based segmentation technique and dynamic programming. The model is based on information obtained from HTTP requests generated by client to a web server. We have evaluated our method on CSIC 2010 HTTP Dataset achieving satisfactory results.

Keywords: *Cyber attacks detection, cyber security, application layer, anomaly detection.*

1. INTRODUCTION

Blameless Bayes and Principal Component Analysis (PCA) were been used with the KDD99 dataset by Almansob and Lomte.Similarly, PCA, SVM, and KDD99 were used Chithik and Rabbani for IDS. In Aljawarneh et al's. Paper, their assessment and examinations were conveyed reliant on the NSL-KDD dataset for their IDS model Composing inspects show that KDD99 dataset is continually used for IDS .There are 41 highlights in KDD99 and it was created in 1999. Consequently, KDD99 is old and doesn't give any data about cutting edge new assault types, example, multi day misuses and so forth. In this manner we utilized a cutting-edge and new CICIDS2017 dataset in our investigation.

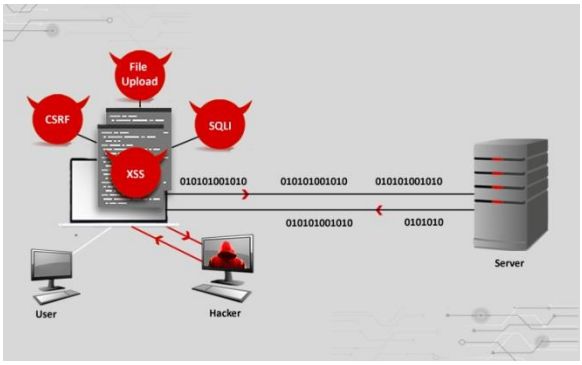

Fig.1: Cyber attacking

Network security is closely related to computers, networks, programs, various data, and so forth, where the purpose of defense is to prevent unauthorized access and modification. However, the growing number of internet-connected systems in finance, E-commerce, and military makes them become targets of network attacks, resulting in large quantity of risk and damage. The aim of project is necessary to provide effective strategies to detect and defend attacks and maintain network security. Furthermore, different kinds of attacks are usually required to be processed in different ways. How to identify different kinds of network attacks thus becomes the main challenge in domain of network security to be solved, especially those attacks never seen before.

Recently the number of security incidents reported all over the world has increased. The national CERTs (e.g. CERT Poland [1]) report that the number of attacks has increased significantly when compared to the previous years. According to the report [1] in 2012 there were 1082 incidents, which is an increase of nearly 80% in comparison to the previous year, mainly due to malware and phishing. The increased number of incidents is strongly related to the increased number of mobile device users who form the population of connect-from-anywhere terminals and regularly test the traditional boundaries of the network security. Also the so-called BYOD (bring your own device [4]) trend exposes the traditional security of many enterprises to novel and emerging threats. Many of nowadays malwares like ZITMO (Zeus In The Mobile) do not aim at mobile device itself but rather on gathering the information about the users, their private data and gaining the access to remote services like banks and web services. There is also a significant number of reported incidents that are connected with a huge widespread adoption of the social media. This trend has an impact on accelerated spread of different kinds of malwares and viruses. As reported by SophosLabs [2] in 2013, botnets have become more widespread, resilient and camouflaged and they are finding some dangerous new targets

2. LITERATURE REVIEW

2.1 Defensive programming: Using an annotation toolkit to build dos-resistant software

ABSTRACT: This paper describes a toolkit to help improve the robustness of code against DoS attacks. We observe that when developing software, programmers primarily focus on functionality. Protecting code from attacks is often considered the responsibility of the OS, firewalls and intrusion detection systems. As a result, many DoS vulnerabilities are not discovered until the system is attacked and the damage is done. Instead of reacting to attacks after the fact, this paper argues that a better solution is to make software defensive by systematically injecting protection mechanisms into the code itself. Our toolkit provides an API that programmers use to annotate their code. At runtime, these annotations serve as both sensors and actuators: watching for resource abuse and taking the appropriate action should abuse be detected. This paper presents the design and implementation of the toolkit, as well as evaluation of its effectiveness with three widely-deployed network services.

2.2 A classification of sql-injection attacks and countermeasures

ABSTRACT: SQL injection attacks pose a serious security threat to Web appli- cations: they allow

attackers to obtain unrestricted access to the databases underlying the applications and to the potentially sensi- tive information these databases contain. Although researchers and practitioners have proposed various methods to address the SQL injection problem, current approaches either fail to address the full scope of the problem or have limitations that prevent their use and adoption. Many researchers and practitioners are familiar with only a subset of the wide range of techniques available to attackers who are trying to take advantage of SQL injection vulnerabilities. As a consequence, many solutions proposed in the literature address only some of the issues related to SQL injection. To address this problem, we present an extensive review of the different types of SQL injection attacks known to date. For each type of attack, we provide descriptions and examples of how attacks of that type could be performed. We also present and analyze existing detection and prevention techniques against SQL injection attacks. For each tech- nique, we discuss its strengths and weaknesses in addressing the entire range of SQL injection attacks.

2.3 SAS: semantics aware signature generation for polymorphic worm detection.

ABSTRACT: String extraction and matching techniques have been widely used in generating signatures for worm detection, but how to generate effective worm signatures in an adversarial environment still remains challenging. For example, attackers can freely manipulate byte distributions within the attack payloads and also can inject well-crafted noisy packets to contaminate the suspicious flow pool. To address these attacks, we propose SAS, a novel Semantics Aware Statistical algorithm for automatic signature generation. When SAS processes packets in a suspicious flow pool, it uses data flow analysis techniques to remove non-critical bytes. We then apply a Hidden Markov Model (HMM) to the refined data to generate state-transition-graph based signatures. To our best knowledge, this is the first work combining semantic analysis with statistical analysis to automatically generate worm signatures. Our experiments show that the proposed technique can accurately detect worms with concise signatures. Moreover, our results indicate that SAS is more robust to the byte distribution changes and noise injection attacks comparing to Polygraph and Hamsa.

2.4 A novel model for detecting application layer DDoS attacks.

ABSTRACT: Countering distributed denial of service (DDoS) attacks is becoming ever more challenging with the vast resources and techniques increasingly available to attackers. DDoS attacks are typically carried out at the network layer. However, there is evidence to suggest that application layer DDoS attacks can be more effective than the traditional ones. In this paper, we consider sophisticated attacks that utilize legitimate application layer HTTP requests from legitimately connected network machines to overwhelm Web server. Since the attack signature of each application layer DDoS is represented in abnormal user behavior, we propose a counter-mechanism based on Web user browsing behavior to protect the servers from these attacks. In contrast to prior works, we explore hidden semi-Markov model to describe the browsing behaviors of Web users and apply it to implement the anomaly detection for the application layer DDoS attacks which simulate the Web request behaviors of browser and use HTTP requests to launch attacks. By conducting an experiment with a real traffic data, the model shows that it is effective in measuring the user

behaviors and detecting the application layer DDoS attacks.

2.5 Robust anomaly detection using support vector machines.

ABSTRACT: Using the 1998 DARPA BSM data set collected at MIT's Lincoln Labs to study intrusion detection systems, the performance of robust support vector machines (RSVMs) was compared with that of conventional support vector machines and nearest neighbor classifiers in separating normal usage profiles from intrusive profiles of computer programs. The results indicate the superiority of RSVMs not only in terms of high intrusion detection accuracy and low false positives but also in terms of their generalization ability in the presence of noise and running time.

3. IMPLEMENTATION

Blameless Bayes and Principal Component Analysis (PCA) were been used with the KDD99 dataset by Almansob and Lomte.Similarly, PCA, SVM, and KDD99 were used Chithik and Rabbani for IDS. In Aljawarneh et al's. Paper, their assessment and examinations were conveyed reliant on the NSL-KDD dataset for their IDS model Composing inspects show that KDD99 dataset is continually used for IDS.There are 41 highlights in KDD99 and it was created in 1999. Consequently, KDD99 is old and doesn't give any data about cutting edge new assault types, example, multi day misuses and so forth. In this manner we utilized a cutting-edge and new CICIDS2017 dataset in our investigation.

Disadvantages:

- In many cases false positives are more frequent than actual threats.
- They don't take care to monitor the false positives, real attacks can slip through or be ignored.

In this paper, a new modified artificial bee colony (ABC), called Mutation Based ABC (MABC) is described. The proposed algorithm give emphasis to process of finding under-utilized servers available in the data centers. For using the resources on cloud, the users use Internet for sending their job request. Cloud Service Provider ensures that each submitted resource is allocated to some VM for execution. For this, the submitted task may move from one data centre to other looking for an under-utilized resource. n turn these data centers may divide the submitted task to sub-tasks commonly known as jobs. The data centers search for the under-utilized resources in the data centers for allocating the jobs to them.

Advantages:

The proposed algorithm is able to minimize the make span time of the jobs by assigning it to the available under-utilized data centers.

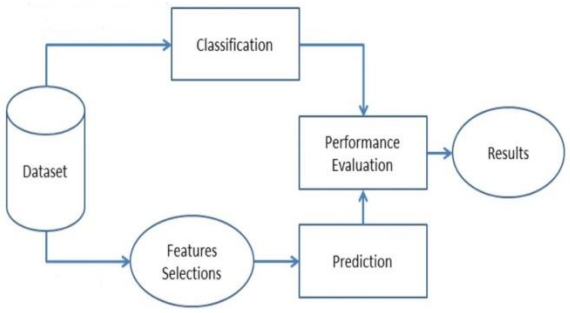

Fig.2: System architecture

MODULES:

- Upload Train Dataset
- Upload Test Dataset

- Preprocess Dataset
- Model Generation
- Run Needleman-Wunsch Dissimilarities
- Training Samples Vs TP Rate

4. MACHINE LEARNING

Before we take a look at the details of various machine learning methods, let's start by looking at what machine learning is, and what it isn't. Machine learning is often categorized as a subfield of artificial intelligence, but I find that categorization can often be misleading at first brush. The study of machine learning certainly arose from research in this context, but in the data science application of machine learning methods, it's more helpful to think of machine learning as a means of building models of data. Fundamentally, machine learning involves building mathematical models to help understand data. "Learning" enters the fray when we give these models tunable parameters that can be adapted to observed data; in this way the program can be considered to be "learning" from the data. Once these models have been fit to previously seen data, they can be used to predict and understand aspects of newly observed data. I'll leave to the reader the more philosophical digression regarding the extent to which this type of mathematical, model-based "learning" is similar to the "learning" exhibited by the human brain.Understanding the problem setting in machine learning is essential to using these tools effectively, and so we will start with some broad categorizations of the types of approaches we'll discuss here.

Categories Of Machine Leaning :-

At the most fundamental level, machine learning can be categorized into two main types: supervised learning and unsupervised learning.

Supervised learning involves somehow modeling the relationship between measured features of data and some label associated with the data; once this model is determined, it can be used to apply labels to new, unknown data. This is further subdivided into classification tasks and regression tasks: in classification, the labels are discrete categories, while in regression, the labels are continuous quantities. We will see examples of both types of supervised learning in the following section.

Unsupervised learning involves modeling the features of a dataset without reference to any label, and is often described as "letting the dataset speak for itself." These models include tasks such as clustering and dimensionality reduction. Clustering algorithms identify distinct groups of data, while dimensionality reduction algorithms search for more succinct representations of the data. We will see examples of both types of unsupervised learning in the following section.

Applications of Machines Learning :-

Machine Learning is the most rapidly growing technology and according to researchers we are in the golden year of AI and ML. It is used to solve many real-world complex problems which cannot be solved with traditional approach. Following are some real-world applications of ML −

- Emotion analysis
- Sentiment analysis
- Error detection and prevention
- Weather forecasting and prediction

- Stock market analysis and forecasting
- Speech synthesis
- Speech recognition
- Customer segmentation
- Object recognition
- Fraud detection
- Fraud prevention
- Recommendation of products to customer in online shopping.

5. EXPERIMENTAL RESULTS

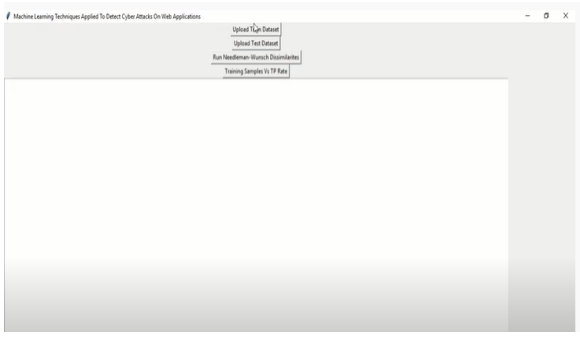

Fig.3: Home screen

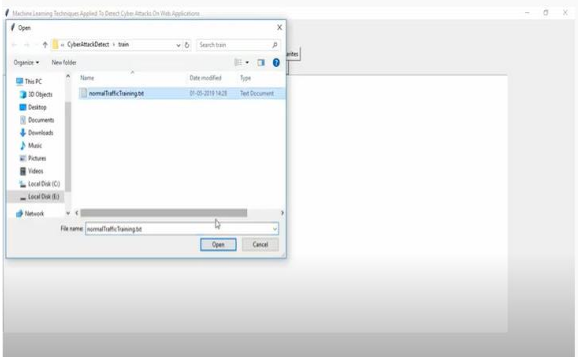

Fig.4: Uploading screen

Fig.5: Train data

Fig.6: Test data

Fig.7: Run Needleman-Wunsch Dissimilarities

Fig.8:Detection result

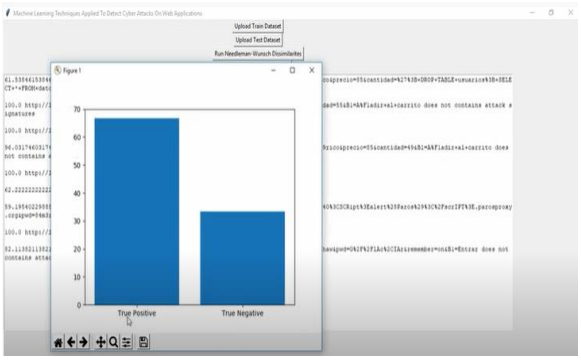

Fig.9: graph

6. CONCLUSION

In this article, the method for application layer attack detection based on machine learning was proposed. The model consists of patterns (in form of PCRE regular expressions) that are obtained using graph-based segmentation technique and dynamic programming. The regular expressions are used for modelling the genuine behaviour of the applications and detecting cyber attacks. We also presented the results that prove the efficiency of the proposed algorithm that can be effectively used for application layer attack detection. The experiments on CSIC'10 show that the proposed approach can achieve 94.46% of detection ratio while having <4.5% of false positives.

7. FUTURE SCOPE

Finally, multiple researchers intend in their future work to convert the models they built into a real-time system in order to benefit from them in real-life scenarios such as in attack detection and prevention. There are two levels of real-time ML which are online predictions and online learning. Online prediction means making predictions in real-time. Furthermore, online learning allows for the system to incorporate new data and update the model in real-time. Hence, converting intelligent models into real time systems may be considered as a fundamental direction to probe by more researchers.

REFERENCES

[1] Thailand Motor Vehicle Registered. CEIC Data Global Database.

[2] Linda, S., Can public transport compete with the private car? Iatss Research, 2003. 27(2): p. 27-35.

[3] CO2 emissions (metric tons per capita) - Thailand. THE WORLD BANK.

[4] Cohen, A.J., et al., Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. The Lancet, 2017. 389(10082): p. 1907-1918.

[5] Litman, T. and D. Burwell, Issues in sustainable transportation. International Journal of Global Environmental Issues, 2006. 6(4): p. 331-347.

[6] Liu, M. and N. Choosri.. A technical solution to improve the red cab for touring in Chiang Mai: Chinese tourists' perspective. in 2016 Chinese Control and Decision Conference (CCDC). 2016. IEEE

[7] Farooq, M.U., A. Shakoor, and A.B. Siddique. GPS based Public Transport Arrival Time Prediction. in 2017 International Conference on Frontiers of Information Technology (FIT). 2017. IEEE.

[8] Bin, Y., Y. Zhongzhen, and Y. Baozhen, Bus arrival time prediction using support vector machines. Journal of Intelligent Transportation Systems, 2006. 10(4): p. 151-158.

[9] Maiti, S., et al. Historical data based real time prediction of vehicle arrival time. in 17th International IEEE Conference on Intelligent Transportation Systems (ITSC). 2014. IEEE.

[10] Fan, W. and Z. Gurmu, Dynamic travel time prediction models for buses using only GPS data. International Journal of Transportation Science and Technology, 2015. 4(4): p. 353-366.