Machine Learning BasedIntrusion Detection System in Cloud Environment

Muhammad Salman Saeed, MS[1]; Raman Saurabh, MS[1]; Sarang Balasaheb Bhasme, MS[1]; Alexey Nazarov, Professor,[1,2]

[1]Moscow Institute of Physics and Technology, [2]Federal Research Center Computer Science and Control of RAS

Abstract

Abstract— Attacks and intrusions on computer networks often have different characteristics and behaviors that require professional help. The number of attacks is growing in line with the development of computer networks. In fact, expert knowledge is declining over time and should be reviewed and made available in the system, making it necessary to hire an experienced person. Cloud computing is an advancement in IT technologies that provide users with the most up-to date,highly sought-after virtual services with high flexibility, low infrastructure costs, and minimal maintenance. Protection against network intruders is one of the most important security challenges for cloud computing, as it affects the privacy, availability, and integrity of cloud services and services. Because cloud computing is a shared environment, it can be vulnerable to various threats. Because Building strong access systems is essential to maintaining cloud security, but it remains an obscure goal and a major challenge due to the growing number of cyber-attacks. In this study, a machine based approach was used to propose an intrusion detection system (IDS) based on a network based cloud computing model. During the previous processing phase, we used a feature selection algorithm from the CICDDoS2019 database to select features. We have used the Naive Bayes Classifier, Decision Tree Classifier, Supporting Classifier, Logistic Regression, and Random Forest Classifier, all of which are well-known categories. Simulation results when compared to these five different Classifiers and another method using a random forest algorithm show improvements in overall accuracy, precision, recall and f1 score.

Keywords— cloud computing, network intruders, intrusion detection, feature Selection, Random Forest Classifier

Introduction

Cloud computing is provided to gain access to storage services and computing resources of users. In fact, the cloud environments with better resource utilization help users to reduce them the cost of access to services. The National Institute for Standards and Technology (NIST) has assuming the following definition of cloud computing [1]: cloud computing is a model for enable ubiquitous, convenient, on demand network access to a shared pool of configurable networks computing resources (e.g., networks, servers, storage, applications, and services) that can be deployed and shared quickly with minimal administrative overhead or interaction with service providers. Key cloud services include: Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). Computer and storage resources are available in cloud environments virtually and dynamically among a large number of users according to their needs and assigned Use. With the growing number of cyber-attacks there is an enormous one requirement for effective intrusion detection systems (IDS) for cloud environment. As the Invasion becomes complicated and difficult to detect, better techniques are used to maintain trust and security in the network. Many in the past ten years Methods are designed to provide reliability, privacy, and security to users Information security. Traditional intrusion detection and prevention solutions, such as firewalls, access control mechanisms, and encryptions, have significant flaws when it comes to fully defending networks and systems from more complex attacks such as denial of service. Furthermore, most systems based on these techniques have a high rate of false negatives detection, as well as a lack of ability to adapt to changing malevolent behaviors. Several Machine Learning (ML) techniques have, however, been applied to the challenge of intrusion detection in the last decade in the hopes of boosting detection rates and adaptability. Intrusion detection is the process of dynamically monitoring events that occur in a Computer system or network, analyzes them for signs of possible incidents and often prohibition of unauthorized access. This is usually achieved by automatically collecting information from a Cloud sources Analyze the information for possible security problems. The development of IDS systems for the cloud environment requires special properties of these systems must be taken into account. Some of these characteristics include: performance, Scalability, diversity of users and services, and dynamics of services. The creation of IDS systems for the cloud environment necessitates taking into consideration unique aspects of these systems. Performance, scalability, diversity of users and services, and dynamicity of services are some of these features. In Cloud [2], there are primarily four types of IDS: Host based Intrusion Detection System (HIDS) Network based Intrusion detection System (NIDS) Hypervisor based Intrusion Detection System Distributed Intrusion Detection System (DIDS) Host Intrusion Detection System (HIDS) detects intrusions using information received from a specific host it is monitoring. The file system utilized, the termination of network activities, and so on are all examples of information. Changes in the host file system and application behavior are also monitored. It will report an attack if abnormal activity is detected. This technique has the disadvantage of being vulnerable to malware attacks. Network intrusion detection systems (NIDS) detect intrusions through network connections and from outside the host machines. This strategy is less vulnerable to malware attacks. Its difficulty stems from a lack of visibility into the host machine's internal state. The cloud provider is responsible for the installation of NIDS in a cloud system. The intrusion detection system in hypervisor-based IDSs runs at the hypervisor layer. It enables users to track and analyze communications between virtual machines, between hypervisors, and within the hypervisor-based virtual network. One of the advantages of hypervisor-based IDS is the availability of information. A Distributed IDS (DIDS) is made up of numerous IDS (e.g., HIDS, NIDS, etc.) spread throughout a vast network, all of which communicate with one another or with a central server for network monitoring. The intrusion detection components gather system data and convert it into a standardized format that can be sent to the central analyzer. A central analyzer is a machine that collects data from multiple IDS and analyzes it. For the analysis, a combination of anomaly and signature-based detection algorithms are applied. Because it combines the benefits of both NIDS and HIDS, DIDS may be used to identify both known and unknown assaults. We focus on a network-based model for usage in a cloud context in this study. The Naive Bayes Classifier, Decision Tree Classifier, Supporting Classifier, Logistic Regression, and Random Forest Classifier, were utilized to detect anomalous behavior. We anticipate that Random Forest Classifier technique will give improved cloud attack detection. The main contributions of this paper can be summarized as follows: • We propose a faster and accurate DDoS detection attacks using a Random Forest Classifier. • We applied our approach with five Machine-Learning algorithms and detect which technique gives high accuracy. The structure of the paper is laid out as follows: The related work is described in section II. We described the different methods which can apply for building an intrusion detection system in section III. The suggested architecture is detailed in depth in section IV. Section V gives the simulation results, and Section VI concludes with conclusion and recommendations for future research.

Methods and Materials

Despite the rapid growth in popularity of cloud services, assuring the security and availability of data, resources, and services is still a major problem. DDoS (distributed denial of service) assaults are not a new threat; they are a serious security concern and a hot focus of research. We'll go through the numerous DDoS intend and launch methods that could be used to conduct or facilitate DDoS assaults, as well as intrusion detection methodologies and defense tactics.

a.DDoS Attack:

DDoS assaults are meant to prevent genuine users from accessing a certain network resource. The Open Systems Interconnection Model (OSI model) can help us comprehend the many forms of DDoS attacks we face. DDoS attacks are designed to attack specific layers of a network connection (application layer attacks target layer 7, protocol layer attacks target layers 3 and 4). There are now two basic ways to initiate a DDoS assault on the Internet. The initial step is to send the victim some faulty packets (i.e., vulnerability attack) or flooding attacks.

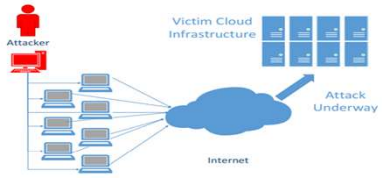


figure 1

An attacker using the second method tries to do one or both of the following [14]: By wasting bandwidth, router processing capacity, or network resources, you can break a valid user's connection. These are essentially flooding assaults at the network/transport level. (For example, flooding attacks)
Make server resources exhausted (e.g., sockets, CPU, RAM, disk/database bandwidth, and I/O bandwidth) to disrupt legitimate users' services. Application-level flooding attacks are a good example of this.
The motivations of the attacker: DDoS attackers are frequently motivated by a variety of factors. Analysis of the attacker's motivations aids in the prevention and response to these attacks [14]. Economic/financial gain, Intellectual Challenge, Cyber warfare.

b. Methodologies for Intrusion Detection:

Signature-based Detection (SD), Anomaly-based Detection (AD), and Stateful Protocol Analysis (SPA) are the three basic types of these approaches. Following that, we'll go over the advantages and disadvantages of the three detection methods.
Signature-based Intrusion Detection: Signature-based Intrusion Detection systems detect intrusions by observing events and recognizing patterns that match known attack signatures. The key events required to carry out the attacks, as well as the order in which they must be carried out, are defined by an attack signature.
Furthermore, it only detects assaults whose signatures have been previously saved in the database. Signatures must be updated on a frequent basis for efficiency. New threats against your hosts are discovered on a regular basis, just as new threats are released on a regular basis, necessitating signature changes. Only the fixed behavioral pattern is effective with this strategy. They are incapable of dealing with attacks launched by humans or worms with self-modifying behavioral traits.
Anomaly-based Detection, Stateful Protocol Analysis (SPA)

c. Machine Learning Techniques:

The various machine learning algorithms and the problem domains in which they are commonly applied are briefly described in this section. In the literature, many decision tree and rule induction techniques have been proposed. The Naive Bayes algorithm is a probabilistic classifier that assumes a variable's effect on a given class is independent of the values of other variables. Class conditional independence is the term for this assumption. One of the most well-known and often used categorization methods is the decision tree.
The divide-and-conquer approach is used in its data structure. Internal decision nodes and leaf endpoints make up the decision tree. The fm (x) function is implemented using discrete outcomes in each decision node, m, to denote branches based on the function's output.
The SVM technique is gaining popularity due to its capacity to generalize, particularly when there are a large number of features m and a small number of data points n. SVM training of a dimensional quadratic programming (QP) issue, on the other hand, requires huge matrix operations, which result in a large number of computations, resulting in poor performance. For decades, the SVM method has been utilized for both anomaly identification and misuse detection. Because of its accuracy and performance, the SMO algorithm was utilized in this work to construct an intrusion detection system.
Leo Breiman created the Random Forests algorithm, which is "a combination of tree predictors in which each tree is reliant on the values of a random vector generated independently and with the same distribution for all trees in the forest." In standard trees, each node is partitioned based on the best partitioning of all variables. In random forests, on the other hand, each node is split differently, with the best partitioning among the subset of predictors that is randomly selected at this node; this helps to avoid the overfitting problem. In addition to its ability to process high-dimensional data, this is one of the key advantages of the random forest algorithm. Logistic regression is a machine learning classification technique. The dependent variable is modeled using a logistic function. The dependent variable is dichotomous, which means that only two classes are conceivable. As a result, while working with binary data, this strategy is applied.

Proposed Work

The random forest classifier is utilized to select features from the dataset in this suggested work. The Random Forest Classifier was chosen because of the tree-based techniques employed by Random Forests, which, of course, improve the node cleanliness. This means that pollution will be reduced across the board (called Gini pollution). The pollution-reduction nodes are found at the beginning of the trees, whereas the pollution-reduction nodes are found at the end of the trees. We can build a subset of the most critical functions by pruning trees below a specific node.

Figure 2 depicts the suggested model's flowchart, which will find improved accuracy within machine learning algorithms. Authenticate and validate features will be sent on to the fit RFC on the training set after the training data has gone through the feature extraction stage. It will be filtered by the train applied machine learning models and v-cross-validation before being sent to extract essential features. After these two stages have been corrected, the data will be run through the evaluation model, and accurate data will be returned.

Figure 2

According to the equation of their function, this proposed model has accuracy, precision, and recall,
where T: true, F: false, N: negative, and P: positive

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

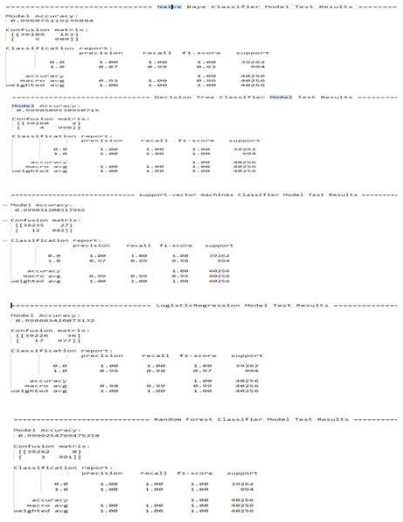
$$\text{Recall} = \frac{TP}{TP+F} \quad (3)$$

$$\text{F1 – Score} = 2 * \frac{\text{Precision}*\text{Recall}}{2*\text{Precision}+\text{Reca}} \quad (4)$$

Using the above explained architecture to achieve high accuracy with the Port Map (CICDDoS2019) dataset, the Random Forest model achieves a better outcome as compared to other tested models.

Results

Using the Random Forest architecture to build a more secure and accurate IDS, we were able to achieve the greatest accuracy result, according to the Port Map Dataset.



We were become successful in achieving the accuracy of 99.99% by using random forest classifier which is almost equal to the perfect system. Random forest architecture is the most suitable architecture for the detection of anomaly behavior among the all above applied architectures.

Conclusions

In this paper, more than one type of machine learning based IDS are applied for finding which one gives the most perfect accuracy results. According to our applied machine learning algorithms Random forest classifier gives the accuracy of 99.99% greater than all the other four applied algorithms on the dataset of Port Map (CICDDoS2019) dataset. The finding results is almost equal to the accuracy of 100% that can boost the benefits of using better models to improve the authenticity and accuracy of detecting intruder attacks. Finally, we want to construct a prototype system based on the suggested mechanism for real-time attack traffic detection and mitigation in order to address security concerns in the future and Models with alternative combinations of machine learning algorithms could be built in the future to obtain greater performance, and IDS with two or more machine learning algorithms could be developed and evaluated on different cloud environments.

Contact

Muhammad Salman Saeed
Moscow Institute of Physics and Technology
Email: said.ms@phystech.edu
Phone: +7 930 933-34-94

References

[1] The NIST Definition of Cloud Computing (NIST Special Publication 800-145).
[2] Modi, C., Patel, D., Patel, H., Borisaniya, B., Patel, A. and Rajarajan, M. (2013). A survey of intrusion detection techniques in Cloud. Journal of Network and Computer Applications, 36(1), pp. 42-57. doi: 10.1016/j.jnca.2012.05.003.
[3] Changsheng Xiang , Zhou Yu , Xilong Qu , "SupportVectorMachine Optimized by Improved Genetic Algorithm" Telkomnika Indonesian Journal of Electrical Engineering 2014.
[4] Changsheng Xiang , Yong Xiao , Peixin Qu , Xilong Qu , "Network Intrusion Detection Based on PSO-SVM", Telkomnika Indonesian Journal of Electrical Engineering 2014.
[5] Alma Husagic-Selman , Rasit Koker ,Suvad Selman, "Intrusion Detection using Neural Network Committee Machine", International Conference on Information, Communication and Automation Technologies (ICAT) 2013.
[6] Daniel Grzonka , Agnieszka Jakobik , Joanna Kołodziej , Sabri Pllana , "Using a multi-agent system and artificial intelligence for monitoring and improving the cloud performance and security", Elsevier 2017.
[7] S. Pal, S. Khatua, N. Chaki, and S. Sanyal, "A New Trusted and Collaborative Agent Based Approach for Ensuring Cloud Security", International Journal of Engineering 2012.
[8] Zhe Li, Weiqing Sun, Lingfeng Wang, "A Neural Network Based Distributed Intrusion Detection System On Cloud Platform", IEEE 2012.
[9] Preeti Mishra, Emmanuel S. Pilli , Vijay Varadharajan, Udaya Tupakula, "NvCloudDS: A Security Architecture to Detect Intrusions at Network and Virtualization Layer in Cloud Environment", Conference on Advances in Computing, Communications and Informatics 2016.
[10] Preeti Mishra, Emmanuel S. Pilli, Vijay Varadharajan,Udaya Tupakula, "Efficient Approaches for Intrusion Detection in Cloud Environment", International Conference on Computing, Communication and Automation (ICCCA2016).
[11] Kleber, schulter, "Intrusion Detection for Grid and Cloud Computing", (EEE Journal: IT Professional, 19 July 2010.
[12] Chirag Modi, Dhiren Patel, Bhavesh Borisanya, Avi Patel, and Muttukrishnan Rajarajan. A novel framework for intrusion detection in cloud. In Proceedings of the Fifth International Conference on Security of Information and Networks, pages 67–74. ACM, 2012.
[13] K.Deepa, and M. Chatterjee. "An adaptive distributed intrusion detection system for cloud computing framework." In Recent Trends in Computer Networks and Distributed Systems Security, pp 466-473, Springer, Berlin, Heidelberg, 2012.
[14] Marwane Zekri, Said El Kafhali, Noureddine Aboutabit and Youssef Saadi. "DDoS attack detection using machine learning techniques in cloud computing environments." Conference: 3rd International Conference of Cloud Computing Technologies and Applications (CloudTech), 2017.
[15] Amar Amouri, Vishwa T. Alaparthy and Salvatore D. Morgera. "A Machine Learning Based Intrusion Detection System for Mobile Internet of Things." Advanced Intrusion Detection & Mitigation Systems in Wireless Sensor Networks, Sensors 2020, 20(2), 461.