Machine Learning Based Hybrid Intrusion Detection For Virtualized Infrastructures In Cloud Computing Environments

Ayesha Sarosh
MBA Scholar, Ottawa university, 4370 W 109th St #200, Overland Park, KS 66211, United States

Abstract. Nowadays technology steady shift have seen from the models of conventional software to the cloud technologies. Cloud computing is rapidly becoming the standard by fulfilling the computer infrastructure demands of all sizes of enterprises. One of the essential tools forbuilding trustworthy & secure environment of Cloud computing is the Intrusion detection, given the ubiquitous cyber attacks which can proliferate morph dynamically & rapidly. Machine learning (ML) based hybrid intrusion detection for virtualized infrastructures in cloud computing environments is presented in this paper. This infrastructure uses Hybrid algorithm: SVM (support vector machine) & K-means clustering classification algorithm, for improving the anomaly detection system accuracy. For evaluating this approach, UNSW-NB15 study is utilized from dataset & results compared with earlier techniques. For evaluating the performance of suggested technique utilizes performance measures like average detection time. This approach has better accuracy compared to earlier approaches.

Keywords: Cloud Intrusion Detection, Machine Learning, Support Vector Machine (SVM), K-means.

1. Introduction
Rapidly cloud computing becomes the standard by fulfilling the computer infrastructure demands of all sizes of enterprises [1]. The cloud computing is based on resource sharing for attaining scalable & coherence economies like a public utility [2]. In the scientific domain utilized non-commercial applications are modeled as work flow applications which can demand numerous computational resources& cloud platform efficiently fulfills these demands.

Failure handling is the most prominent feature of the cloud distributed computing platforms [3]. When an application is running on client side then the cloud is responsible for storing the data of clients. Suppose the data are damaged or lost then it can lead to severe calamity. Therefore in cloud application they hold a very critical position. Data lost cause a critical situation to the enterprises which may leads to business loss & revenue loss. In cloud platform to execute scientific applications, most importantly the data should be recovered within minimal time & ensure minimal damage.

The VM (virtual machines) collection & a monitor machine namely hypervisor, are referred by Virtualized environment [4]. Nowadays the virtualized infrastructures became most popular with public cloud advents like Azure from Microsoft, AWS (Amazon web services) from Amazon [5]. Cyber attacks are also prone by virtualized infrastructure like other technologies. In virtualized environments to detect the cyber attacks IDS (intrusion detection system) is utilized [6].
Execution of a task in cloud platform may result in failure due to different reasons like service level agreement (SLA) violation over resources, resources over cost & VMs over usage, etc. In cloud platform to deploy the scientific workflow models requires task failures minimization & fault tolerance improvement. A cloud service allows a user to access & share their system often on the internet with minimal managing efforts. It can offer the user for minimizing the cost of up-front IT infrastructure and maintenance cost. In a cloud the user desired to implement his/her system by using a IaaS (infrastructure as a service) [7]. Since the cloud provider allows IT infrastructure & the user does not have to configure it.

For analyzing the large amount of data produced from transactions between the cloud computing device systems requires sophisticated & modern technologies because of distributed infrastructure size & complexity of cloud computing networks [8]. Generated complex data should be analyzed in real-time for developing robust IDS. Hence ML techniques like clustering & classification are promising in cloud networks to analyze the big data containing traffic flows for cues of intrusion even when they have signatures which are not encountered before [9]. ML techniques confirm attacks possibility over the features of collected log for classifying the paths as benign or malicious.

For virtualized infrastructure an adaptive hybrid network-based IDS model is presented here to determine the abnormal behavior of a network in cloud computing domain. This project main contribution is:

A general survey over the problem of anomaly detection is performed in present studies.

A hybrid ML technique for network intrusions effective detection, IDS is suggested for overcoming the shortcomings existed.

2. Virtualized Infrastructures in Cloud Computing Environment

However, IDS are various kinds; one of most effective IDS is anomaly based systems [10]. Therefore anomaly IDS are suggested for many ML based techniques. ML techniques try to model intrusion detection as a problem classification.

2.1 Virtual Machine Security

Virtual machines security can be provided by monitoring the memory behavioral patterns & system calls is a challenging task if the attacks are easily escapes the static analysis approaches as the memory sequence is dynamically modified [11]. The attacks may also have chances for adding irrelevant memory system calls to modify the memory sequence for escaping the static detection techniques. Dynamic detection techniques are better for attacks on VM. Significant improvement is required for VM attack detection based on behavioral patterns because many of detection techniques doesn’t focus over attack patterns intrinsic behavior. Protection of VMs by behavioral patterns is the major setback. Interaction between the resources of corresponding operating system & system calls is monitored by malware tool namely Accessminer.

2.2 Anomaly Detection in Cloud Computing

For determining the abnormalities in the dynamic cloud infrastructures, utilizing ML approaches is defined by anomaly detection [12]. In virtual machines for verifying the deviations in general work it can make statistical analysis usage. For project identification, prediction & detection and behavioral analysis anomaly detection technique is extensively utilized [13]. Anomalies corresponded to the system performance are discussed and the root causes in the system which can create bottle necks. Overview of various malware detection approaches according to the system performance is given by the researchers. They state that in cloud infrastructure for detecting the anomalous behaviors is to monitor the VMs execution & obtained the cloud data runtime performance. For refining the future detections they have followed a recursive learning technique.
2.3 Machine learning techniques
Mainly ML techniques are categorized as 2 algorithms: supervised & unsupervised learning. With signature - based technique for building the IDS, utilizes supervised learning algorithms, where signatures of previously known attacks are utilized as training data [14]. For building the signature - based IDS, most widely utilized supervised algorithms are SVM, decision trees, linear regression, linear discriminant analysis, neural networks, logistic regression & Naïve Bayes [15]. According to known signatures behavior of the supervised learning algorithms is adjusted and also known as event or label classifier, for achieving better detection. The signature - based ML IDS accuracy depends on the training dataset quality & the utilized parameters in supervised ML technique. The anomaly - based detection is IDS second type. Based on the normal behavior of network a model is formed for normal activities in this method and any type of deviation in the normal behavior of a network is considered as an attack.

3. ML Based Hybrid Intrusion Detection System
For virtualized infrastructure in cloud computing environment a framework of hybrid IDS based on ML technique is suggested and is shown in Figure 1. Hybrid techniques perform better than normal approaches in intrusion detection systems. So this works mainly focus on the analysis of contextual relationship between the data flow from the network utilizing hybrid techniques. By matching the network or node activity source & destination this approach can collect the networks, application logs and also tries to correlate them. If these correlations are derived then from this correlations potential attack paths are identified through the Map reduce approach. To confirm the attack presence ML techniques are utilized. In detail, for producing the labels automatically utilizes K – means clustering algorithm and for build a learning model SVM is used, that can be utilized to evaluate the new questions.

3.1 Network and user app logs
Collecting the features of an attack is the first step in this method. Utilizes the TShark for obtaining the network logs contain the guestVMs traffic flow. Along with port numbers source & destination IPs are collected. Further from individual VMs the application logs of a user are collected. For obtaining the form correlation the network logs are assembled with application logs: App guest, h IP source IP destination, Port source Port destination, User ID App i.

3.2 Map Reduce technique
Usually, in application logs as well as in networks a cyber attack leaves its foot prints. Therefore the indication of an attack can be more than the normal number of communication logs from a specific source to destination. Map Reduce method is utilized for efficient sift over the generated numerous logs. For determining the occurrence frequency the Map Reduce method runs on the correlation logs. Among these paths one who has greater occurrence count is considered as a threat.

3.3 Data pre-processing
To improve accuracy of IDS a pre - processing task is necessary. So that the traffic analysis consisting 2 crucial concepts; network flow analysis, classification of traffic. Network flow means a group of packets between 2 end points which are captured at hypervisor level. Network management & monitoring tools are required for performing the flow analysis of a network. The output dataset consist some of missing data after monitoring the features extraction & network flow. Primarily focus on the removal of inconsistent data & noisy.

3.4 Dimensionality Reduction
Few features don’t help for the process of detection and might increases the complexity of detection process. So, removal of features which can have a little variance may even cause better results. Since
these dimensions mostly contain noise. Hence reducing the feature weights might have more information. For removal of features PCA (principal component analysis) is utilized. Hence for every time only a component subset of feature vector can be utilized as input for detection algorithm. Depends on the exploratory experiments over the training dataset from the UNSW-NB15, the first 10 features of 16 obtained are decided by PCA to retain. By utilizing PCA the 47 features are reduced to 10 features.

Fig. 1: Hybrid Intrusion Detection System Using Ml
3.5 Clustering and Labeling the Data-set

The suggested detection module contains 2 models; first model is unsupervised clustering for creating relevant data classifiers to the whole data-set. The first model output is applied as input to the supervised model; anomaly detection model build by SVM is represented in Figure 1. It is proved that data label is not presented in network packets. K-means clustering algorithm is employed for determining the abnormal behavior of the network and it can be labeled as an attack, however all the network packets are labeled as normal. 1 is treated as abnormal behavior & 1 is normal behavior. In such cases 3 threshold functions are required, which are based on 3 parameters (i)sparse or not, (ii)small or not & (iii)large or not. Next for determining the clusters density & size 3 threshold parameters are defined.

Definition 1: Threshold functions of small clusters: Assume that the data-set have N data instances & consists K clusters in cluster set C = {C1, C2, C3, ..., Ck}. Any cluster average size is calculated as N/K. If the cluster size is $\frac{i}{\alpha} = \frac{n}{k}$ then it is small. Where $C_i$ is the number of data instances in the cluster $i$ and $0 \leq \alpha \leq 1$.

Definition 2: Threshold function of Large clusters: the only difference between this definition & previous is if cluster size $i = \beta n/k \leq 1$ then the cluster is larger.

Definition 3: The cluster density threshold function: for determining the cluster whether it is a sparse or not. Every cluster consist a centroid position due to the usage of K-means clustering. Assume that the centroid position is $\mathcal{C}$, data point in the cluster is the number of data points in the cluster $d_i$ & the number of data points in the cluster is $M$. Sum of squares average is denoted as $E = 1/M \sum_{i=0}^{M} (\mathcal{C} - d_i)^2$.

If the E value is small, which means high density & vice versa. At this time another parameter $\gamma$ is defined for defining the cluster sparsity level. Thus, $E_i \geq \gamma \times \text{median (E)}$, considered as $i$ cluster is sparse.

3.6 Detection Model

Supervised learning approaches are faster in testing mode compared to unsupervised techniques; SVM is one of the better supervised models are chosen for the ML based detection model. To create a learning model data-set is feed with new label for SVM with polynomial kernel. For further anomaly detection this learned model can be utilized with network new traffic.

4. Evaluation of Hybrid-Detection model

The public cloud of AWS is utilized as cloud environment. For creating the virtual machines the AWS EC2 (Elastic Compute Cloud) is utilized (instances are identified in AWS). As virtual machines Linux instances are utilized. A malware sample is run over one of the VMs during a malware attack for obtaining application logs & network. It is run on a controlled domain to avoid the malware adverse effects. The IDS system is implemented & tested with the collected samples from the data set UNSW-NB15. This system should correctly detect all the types of malwares in real-time.

4.1 Clustering Module

For measuring the impact of applied k-means clustering result is the main purpose of this evaluation which is utilized for creating relevant labels over the data set. Before measuring the clustering algorithm result, describe the critical parameters which are adjusted during the experiments. For creating the relevant labels on data set the parameters ($\alpha$, $\lambda$, $\beta$) are utilized in detection technique. These parameters values are between 0 & 1. For selecting the small clusters in K – mean clustering algorithm result, specifically $\alpha$ parameter value must be close to 0. For selecting the large clusters in K
– mean clustering algorithm result, \( \lambda \) parameter value must be close to 1. For selecting the sparsest clusters in K – mean clustering algorithm result, \( \beta \) parameter value must be close to 1. The generic parameter for K- mean cluster is the number of clusters that are required to adjust for evaluating this approach. During experiment each of these parameter are applied with various values; \( 0.1 \leq \alpha \leq 0.2, \ 0.85 \leq \lambda \leq 0.99 \ & \ 0.85 \leq \beta \leq 0.99 \). These values are utilized in the same model at different nodes & each model result will be compared for selecting the best values of these parameters.

4.2 Anomaly Detection Model

After collecting the labeled data sets from previous stage, SVM model is trained by this labeled data set. Then this trained model is utilized in this detection method to detect new coming data. For these events the SVM algorithm requires a data set which can contain events & a classifier. Usually the classifier label is a binary classifier. In this case, the binary classifier has 2 values, -1 for abnormal activities & 1 for normal activities. 85% of data sets are used for training set & 15% of data sets are used for testing. The system does not raise any alarms when normal applications are running. Therefore false positive cases are eliminated. The system parameters like accuracy, precision & recall are denoted as:

\[
\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)}
\]

\[
\text{Precision} = \frac{TP}{(TP + FP)}
\]

\[
\text{Recall} = \frac{TP}{(TP + FN)}
\]

| ML models                  | Accuracy | Precision | Recall  |
|----------------------------|----------|-----------|---------|
| Random tree                | 73.1     | 70.4      | 71.1    |
| Decision tree              | 81.6     | 76.8      | 80.4    |
| Hybrid model (k-means & SVM)| 89.7     | 82.3      | 85.6    |

**Table 1: Performance Evaluation**

To calculate the average detection time, dividing the sum of detection times for all the samples analyzed by total number of samples.

\[
\text{Avg. Detection time} = \frac{\sum \text{Detection time for individual malware samples}}{\text{Total number of samples analysed}}
\]
The presenting mode is hybrid intrusion model which is combination of K-means and SVM accuracy is better than the other individual intrusion detection systems as shown in Table 1. 70 seconds average detection time is achieved by this system. Correlation of application logs & network logs allowing the user to not only detecting the anomalous network connections but also allows the user for identifying the connection source. System scaling includes multiple VMs, the user logs & network logs are extended for incorporating the virtual machine ID from a specific log entry obtained.

5. Conclusion
In cloud computing environment for detecting the known & unknown attacks, a hybrid IDS, the combination of unsupervised & supervised ML techniques for virtualized infrastructures are suggested in this paper. To develop an effective method for identifying the security analytics & cyber threats in virtualized environments, ML algorithms with big data capabilities are combined. This hybrid IDS is a ML – based technique which is using K – means clustering algorithms for the flow of network clusters. The detection model is the SVM trained model for detecting the anomalies in new events. For evaluating this detection system a publicly available labeled data set UNSW-NB15 is utilized. It is a rich labeled data set, consist nine various types of attacks. This presented IDS model is achieving the accuracy as 89.7% which is better than individual machine learning models as random tree and decision tree. Therefore, the approach successfully identifies the presence of attack in virtualized environments in real time.

References
[1] DivyaPramasaniMohandoss, Jameson Hodge, Kun Suo, Steven Kemp, Yong Shi, “Towards Optimizing Task Scheduling Process in Cloud Environment”, 2021 IEEE 11th Annual Computing and Communication Workshop & Conference (CCWC), Year: 2021
[2] Jun Wen, Wang Zhang, “Billing System in Distributed Computing Environment”, 2020 International Conference on Computer Engineering and Intelligent Control (ICCEIC), Year: 2020
[3] Yamin Thet Htar Hlaing, Tin Tin Yee, “Static Independent Task Scheduling on Virtualized Servers in Cloud Computing Environment”, 2019 International Conference on Advanced Information Technologies (ICAIT), Year: 2019
[4] K. Swedha, Tanuja Dubey, “Analysis of Web Authentication Methods Using Amazon Web Services”, 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Year: 2018
[5] J V Anand Sukumar, I Pranav, MM Neetish, Jayasree Narayanan, “Network Intrusion Detection Using Improved Genetic k-means Algorithm”, 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Year: 2018
[6] Anderson S. Raugust, Felipe R. de Souza, Mauricio A. Pillon, Charles C. Miers, Guilherme P. Koslovski, “Allocation of Virtual Infrastructures on Multiple IaaS Providers with Survivability and Reliability Requirements”, 2018 IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA), Year: 2018.
[7] Tamás Kovácsyázy, “Distributed architecture for real-time cyber-physical system, time-sensitive networks”, 2018 19th International Carpathian Control Conference (ICCCC), Year: 2018
[8] Elif Varol Altay, Bilal Alatas, “Detection of Cyberbullying in Social Networks Using Machine Learning Methods”, 2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT), Year: 2018.
[9] Amol Borkar, Akshay Donode, Anjali Kumari, “A survey on Intrusion Detection System (IDS) and Internal Intrusion Detection and protection system (IIDPS)”, 2017 International Conference on Inventive Computing and Informatics (ICICI), Year: 2017
[10] Ahmed M. Mahfouz, Md Lutfar Rahman, Sajjan G. Shiva, “Secure live virtual machine migration through runtime monitors”, 2017 Tenth International Conference on Contemporary Computing (IC3), Year: 2017
[11] Subama Shakya, Sandeep Sigdel, “An approach to develop a hybrid algorithm based on support vector machine and Naive Bayes for anomaly detection”, 2017 International Conference on Computing, Communication and Automation (ICCCA), Year: 2017
[12] Dandan Ma, Yuan Yuan, Qi Wang, “A sparse dictionary learning method for hyperspectral anomaly detection with capped norm”, 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Year: 2017

[13] Glaris T. Hilda, R. R. Rajalaxmi, “Effective feature selection for supervised learning using genetic algorithm”, 2015 2nd International Conference on Electronics and Communication Systems (ICECS), Year: 2015.

[14] Shu-Xia Lu, JieMeng, Gui-En Cao, “Support vector machine based on a new reduced samples method”, 2010 International Conference on Machine Learning and Cybernetics, Volume: 3, Year: 2010.