Providing Secure Service in The Cloud Environment by Using Machine Learning Algorithms and Using SVD

Bahreini Kambiz[1†] and Arjmandy Khadijeh[2††]

University of Islamic Azad Bushehr Iran, University of Islamic Azad Dubai UAE

Summary
Cloud computing is a new example of service provision. This type of service is a service previously provided by other service providers locally or extensively, such as telecommunication services. Today, one activity, such as processing, storage and other activities, is calculated and paid according to the needs and the amount of services provided. In this situation, because customers do not have any information on how to use data, how to store and how to share data, they are concerned about whether the levels of access are correct and controlled on the data Is done. Cloud computing is now one of the most frequently used innovations in information technology. Most technology companies are currently using or planning to deliver products based on the cloud computing model. In the future, there will be new security threats due to a change in the way services are provided to users and service providers. For this reason, the security, threat, attack, and hack issues should be explicitly addressed in the cloud.
With the approach to the foregoing, this research has been presented to protect the security of the cloud, a model for detecting threats and attacks in the cloud using machine learning techniques.

Key words:
Cloud Computing, Security, Modeling, Optimization, Intrusion Detection.

1. Introduction

Cloud computing is the result of an upgrade in the computational perspective that has emerged on the current and emerging technology. The cloud is considered as a new opportunity for sharing resources and relative innovation in the delivery of services. However, customers at first glance are looking for security in the cloud. Cloud computing requires a new security. As mentioned, cloud computing is an improved sample in the provision of resources and services that can provide important benefits to providers and customers. By virtualizing IT infrastructure and services, customers can create a virtual network and data centers in the least time with minimal cost. Therefore, cloud computing is a software program and service provided in a network distributed by virtual resources through common network protocols. Therefore, clients receive hardware services through Virtualization without pay for buying the hardware, purchase, and support and upgrades software at lower costs. However, due to the fact that in this method of service delivery, network administrators and end users do not have access to the physical location of the resource. Therefore, there should be a specific mechanism to ensure the security of the service provided so that the end-user can use this method with high confidence.

An attack is a dangerous or non-dangerous attack to change or exploit an available resource through the network in a way that was not intended. [1] Unauthorized use of existing resources through the network can lead to unauthorized access to information sources through the network, unauthorized access to network information or attacks that impede the provision of services. The purpose of network security is to protect resources, maintain data integrity and maintain data access against attacks.

Intrusion detection systems are used to monitor malicious and unwanted network traffic. The intrusion detection system uses a traffic information sensor. The sensor connected to the network is downloaded using a traffic data packet collected by the sensor called the Aggregator and then sent to the intrusion analysis system to analyze the traffic data. Aggregator performs data analysis using signature detection or anomaly detection method on the data, and annihilates the result as a natural threat or traffic information. The intrusion detection system administrator can then, based on the information obtained by the system configuration To do. [1] A block diagram of an example of this system is shown in Figure 1.

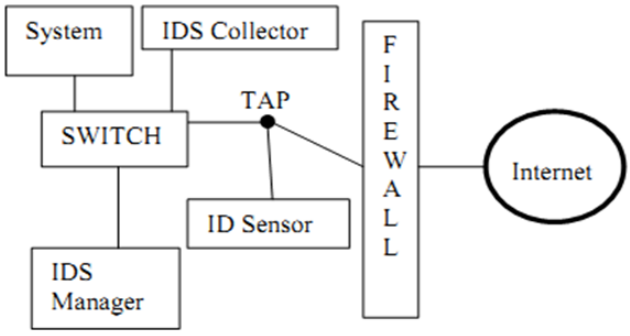

Fig. 1 General architecture of an intrusion detection system.

Manuscript received May 5, 2019
Manuscript revised May 20, 2019

Categories of attack detection system are done from six perspectives. [2]

- From the perspective of attack detection it is include the following: signature detection, anomaly detection.
- From the perspective of protection of the system include the following: attack detection by the host, attack detection by network and coed.
- From the perspective of the structure include the following: centralized system, distributed system.
- From the perspective of data sources: remaining audit effect, network packets, analyzes the mode of system such as the core - services and files.
- From the perspective of system behavior after the attack: active attack detection system, passive attack detection systems.
- From the perspective of education based on the analysis: real-time, non-real-time.

2. Related Researches

In recent years, several work has been done based on the creation of model methods based on data mining. In this method, unknown attacks are often identified by comparing the behavior of known and common attacks. This method is a faster method than manual, which is done by skilled people, because the skilled person must be able to recognize all attacks. Different methods of data extraction are identified for transport identification by [4, 5, and 6]. Each data mining method is presented in order to classify the behavior of the network as a normal state, or attack. Some researchers have used different types of genetic algorithms for Host Intrusion Detection (HIDS) or Network (NIDS) in different scenarios and in different ways [4], resulting in a precision of 0.9500 and a false positive of 30.30. The history of using this algorithm dates back to 1995 when Mr. Crossby and Spfrord were able to detect abnormal behavior of the network with several hosts (agents) and genetic programming. In this scenario, each host is responsible for part of the system's intrusion detection system. If an attack occurs, the detection system works at one of the hosts. In this case, the use of multiple hosts prevents system faults. But this increases the runtime for system updates. SVM was also invented by Vladimir Putin in 1963 and expanded in 1995 by Vapnik and Corinna Cortes for nonlinear mode. SVM (Support Vector Machines - SVMs) is one of the supervised learner methods that can be used for classification and regression. This modeling technique is a relatively new method that has been welcomed by many researchers in recent years. [5, 6 and 7] Amjid Hussein Lackt, using the NB Tree algorithm, was able to conclude the degree of network penetration with a precision of 0.98 and a false positive = 0.1. Also, the Random Forest algorithm is an outline of a recursive tree classification. This method is a good method in terms of the precision of extraction algorithms. Especially for the large data set that has many features. This algorithm generates different trees. Each tree is built with a sample of bootstrap that is different from the initial data. For this purpose, a tree classification algorithm is used. After the formation of the forest, a new specimen needs to be classified. Each tree is classified in the forest. Each tree has one vote, which represents the tree's decision about the class in question. The forest, according to most of the opinions that are given in the object class, determines the class of this object. In the random forest algorithm, there is no need for linear confirmation. Since each tree is made using bootstrap, about one-third of bootstrap samples are left out and not actually used. [6]
Also these algorithms have been used by other various researchers. The results obtained by Ben Amor was measured on Decision Tree algorithm for accurately measure (Accuracy = 92.28%) and on Naïve Bayes algorithm for degree of accuracy (Accuracy = 91.47%).[8]

3. Proposed Approach

In this study, we provided a model based on a machine learning method on virtual machines exist in the cloud. In the proposed model, the machine learning method was performed as follows.

- Using an intrusion detection algorithm based on anomalies.
- Anomaly detection using a combination of several classifieds section.

3.1. Section of virtual machines, servers and service

IDS Service: Intrusion detection service in the network will increase the level of security in the cloud, which from the method perspective has been done in two ways. The first method is based on processing and comparing the user's current behavior with normal behavior in the network. The second method is based on the knowledge base, checking and receiving collected information and analyzing it by artificial intelligence algorithms in order to detect intrusion. Which includes two sub-systems called System Analyzer and System Alert. This service in terms of position, if installed on the Host named HIDS, and if installed between the firewall and the router called NIDS. Both types of IDS in the proposed model have been used in order to provide physical and virtual servers and workstations. In the following, each of these sections of the task in the proposed model is described as follows in Fig. 2

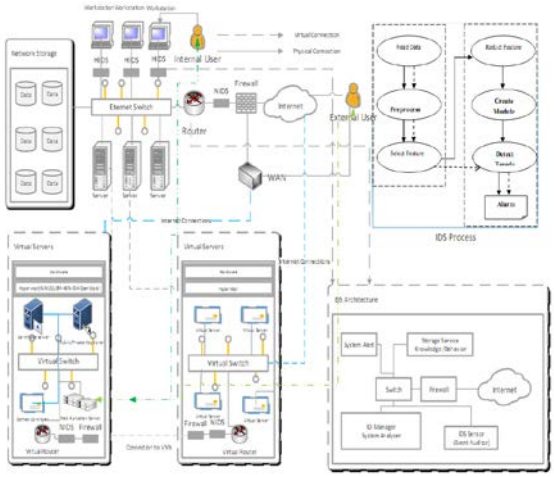

Fig. 2 Model related to Intrusion Detection System by using Virtual Machine Manager in the Cloud

- System Analyzer Task: Attack user on the network has been detected by using normal behavior and comparing it with the abnormal behavior. Also detect attack with analyzes, checks incoming packets and by the rules contained in the database. Then the result is sent to IDS service core and if the calculating probability in order to the presence of attack was high, attacks known and sent a message to other nodes.
- Alert Sub System task: This subsystem when the attack is detected sends attacked nodes information to other nodes that IDS are available onto them.
- Service Storage Services: This storage management service has two databases called the KDD99 Database Knowledge Base and the Behavior Base. When a node receives a request or response, the nodes compare the information with the internal data stored.
- Event Auditor's Task: To identify the threat, the operating system environment, and the status of the data exchanged in the network. This section has the ability to monitor data access by the data analyst.
- Domain Controller: This service is used to provide computer and user authentication services.
- Certificate Server task: This service is created for creating and managing certificates on the basis of Public key Infrastructure (PKI). Through this service you can set out a Certification Authority (CA) Server to create Certificate for secure exchange of Information and data on the internal network. Generally, certificates can be classified in two categories. Self-Sign Certificate which is valid only for the internal network and the other is trusted Certificate which are used in the internal network and the Internet.

3.2. Section of training, testing and validation of intrusion detection model

The model presented in Figure 2 consists of two different sections. The first stage is the training of the system. Then data is converted to numerical data so that it can be learned as input to the device. Using the genetic algorithm, the weight index values are analyzed based on the weight of the information. This is used to reduce the dimensions of the feature. We also use this algorithm to reduce dimensions using the SVD method. This method, using special vectors, tries to reduce the dimension. In some data, some traits are dependent. As the dimming of the model, speeds up, matrix or data sets are selected based on high-weight features. In selecting characteristics from the data set, the selected features should be such that they have more information than the original signal. Finally, the features that weigh more are selected and trained to build the model. It is noteworthy that the use of different methods is used to calculate the weight. In this study, a genetic algorithm is used to calculate weight and reduce traits, and then it is obtained using a machine learning algorithm model, so that the model is produced as an intrusion and diagnostic attack. In order to test the model, the experimental data related to KDD99 have been used and the model has been evaluated. Then calculate the accuracy of the model and response time. In the following figure, the intrusion detection system is provided using a virtual machine management device.

3.3 Used Database

Currently, only database is KDD99 used for intrusion detection database that it has been used in this study. According to that the used model is from monitoring learning methods, KDD99 is the only standard database that has wide attributes in order to training and testing model that used by most of researcher in order to evaluate the model. This database has 24 kinds of attack in training data and 38 kinds of attack in testing data that written as the base in 4 kinds of attack that have been brought in the Table 1.

Table 1: Mappings types of data sets attacks KDD99 to various classifieds attacks in classifieds Section

Class	Cass Attacks in the training data
Probe	Ipsweep,Nmap,Portsweep,Satan
DOS	Back,Land,Neptune,Pod,Smurf,Teardrop
U2R	Buffer overflow,Loadmodule,Perl,Rootkit
R2L	Ftp_write,Guess_passwd,Imap,Multihop,Phf, Spy,Warezclient,Warezmaster

The dataset of 41 attributes of each record includes 7 Attribute of the type of discrete data and 34 is attributing of the type of continuous. Every record is includes various information including service type, protocol type, the sender and receiver IP address, transmitter and receiver port number and the other. This database has 24 kinds of attack in training data and 38 kinds of attack in testing data that are written as a base in 4 kinds of attack that have been brought in Table 2. The dataset of 41 attributes of each record includes 7 Attribute of the type of discrete data and 34 is attributing of the type of continuous. Every record is includes various information including service type, protocol type, the number of unsuccessful attempts to log in, and so forth. According to studies done on this dataset, the following conclusions about the distribution of it data is given according to the Table 2.

Table 2: Dataset Specifications of KDD99

Class	%10 of the data for training	%10 of data for testing	Binary Label
Normal	19.69%	19.48%	normal
Probe	0.8%	1.34%	Anomaly
DOS	79.24%	73.90%	Anomaly
U2R	0.01%	0.07%	Anomaly
R2L	0.23%	5.20%	Anomaly

3.4. The process of providing secure service based on the proposed model

As seen in Figure 2, users are includes 2 types of user in order to use of the Cloud service in the provided model. Users who are enter to internal network via WAN and use of available services in network or users who are in local network and use of existing services via the workstation. For both types of user-for example, if the requested service is a Virtual Web Application Server, the process of providing secure service to users will be in following method.

- Step 1: User enters website address in the browser.
- Step 2: Requested website Information enters the IDS through the pass from Firewall.
- Step 3: IDS Monitor performs incoming Packet via Firewall or Router in order to intrusion detection based on the behavior, or based on intelligent algorithms (Knowledge Base with KNN + SVD).
- • Step 4: Packet situation is investigated in terms of attack. If the attack does not seen, an incoming Packet passed by IDS and transmitted to Step 5.Else if the attack is observed, incoming Packet not pass by IDS and transmitted to Step 7.
- Step 5: User required information and password is sent in order to authentication to, domain controller of virtual servers and certificate.
- Step 6: In case of user authentication in addition to processing of requested service by Virtual Web Application Server the result is displayed to the user as alert. And finally this process is transferred to Step 1.
- Step 7: In case that incoming packet is detected by IDS Sensor and attack IDS Manager, Then System Alert sent attacked nodes information to other nodes which IDS service is enabled on them and preventing from entering packet to network by IDS Manager. Finally, this process is transferred to Step 1.

4. Conclusion

As a result, this research is for achieve the purpose, examine model creation of various methods in order to attack detection in the cloud such as SVM, Random Forest, Auto MLP and KNN that is as Table 3.

Table 3: Results of implementing the different algorithms on KDD99 dataset

Row	Training data	Testing data	Classifier	F.S	Accuracy	Precision	Recall	F-Measure	Time
1	5%	5%	SVM	W>=0.35	82.51	56.01	74.13	63.81	00:36:44
2	10%	10%	KNN	W>=0.3,k=1	99.11	86.03	88.40	87.2	00:16:22
3	10%	10%	KNN+SVD	w>=0,k=1	99.29	95.65	95.26	95.45	00:04:00
4	10%	10%	Auto MLP	w>=0.30	99.54	74.04	77.58	75.77	00:16:23
5	10%	10%	Auto MLP+SVD	W>=0	99.66	76.26	77.37	76.81	00:12:46
6	10%	10%	Random Forest	W>=0	90.71	39.52	36.63	38.02	00:23:53

In this model, the SVD algorithm is used as a precursor to reduce attributes in order to reduce the amount of data and increase the speed of training. To compare rows 2 and 3, it can be seen that the processing time of one quarter and the accuracy level increased to 0.18%, the “Accuracy” value increased to 9.62%, the “Recall” rate increased to 6.86%, the value of “F-Measure” increased to 8.25%. Using the SVD as the pre-processing data, the results of the “Auto MLP” algorithm are similar. The time value is reduced from 16:23 to 12:46 minutes, the “precision” value is increased to 0.12%, the “Accuracy” value is increased to 2.22% The “Recall” rate increased to 0.21% and value of “F-Measure” increased to 1.04%.

In the comparison criterion, using the SVD algorithm, the number of changes in the KNN algorithm is greater than the Auto-MLP, and the random forest algorithm has not been achieved.

However, according to the survey conducted on machine learning algorithms it seems to SVD algorithm could be used as a relatively accurate data pre-processing method in research and mining projects. And also it is recommended that the issues related to security in the cloud in the upper layers of the cloud are also being investigated such as application as services.

In the end, Because of using data set in the cloud in type of attacks in cloud with names R2 and U2L has a low record, This limitation has had a significant impact on the final result. If the number of data of this type of attack in the data set was with other attacks at a level will be cause to increase the accuracy of the model. And also to cause of the volume of data associated with used data set (KDD99) is high, in order to high accuracy must be processed the whole data set data to increase the accuracy of the produced model. To cause of limitations on Rapid miner used version this software is limitation on data volumes and for this reason produced model accuracy will not be the final accuracy. And will be depends on the software and hardware processor.

References

[1] Davari,Dowlat Abadi, Majid, hacker attacks and how to cope, 1391.

[2] Kaveh Pashaei, Ali Abbas, Introduction to Data Mining and Knowledge Discovery old John.

[3] W. Lee, S. J. Stolfo, and K. Mok. Data mining in work flow environments: Experiences in intrusion detection. In Proceedings of the 1999 Conference on Knowledge Discovery and Data Mining (KDD-99), 1999

[4] A. Ghosh and A. Schwartzbard. A study in using neural networks for anomaly and misuse detection. In Proceedings of the Eighth USENIX Security Symposium, 1999.

[5] Cloud Security: Attacks and Current Defenses Gehana Booth, Andrew Soknacki, and Anil Somayaji, 8th ANNUAL SYMPOSIUM ON INFORMATION ASSURANCE (ASIA'13), JUNE 4-5, 2013, ALBANY, NY

[6] Machine Learning Approach for Intrusion Detection on Cloud Virtual Machines International Journal of Application or Innovation in Engineering & Management (IJAIEM) Web Site: www.ijaiem.org Email: editor@ijaiem.org, ditorijaiem@gmail.com Volume 2, Issue 6, June 2013 ISSN 2319 - 4847 Volume 2, Issue 6, June 2013 Page 57

[7] Model Generation for an Intrusion Detection System Using Genetic Algorithms Adhitya Chittur November 27, 2001 Ossining High School Ossining, NY

[8] N. Ben Amor, S. Benferhat and Z. Elouedi. Naive Bayes vs Decision Trees in Intrusion Detection Systems. In Volume 2, Issue 6, June 2013 Page 65 SAC " 04: Proceedings of the 2004 ACM symposium on Applied computing, pages 420-424, New York, NY, USA, 2004. ACM. ISBN 1-58113-812-1.

Kambiz Bahreyni received the B.S. in Electrical Engineering and M.S. degrees in Computer Engineering from Islamic Azad University Bushehr Branch in 1999 and 2015, respectively **.**Since 1999, he has been managing the IT department of management of the social security organization in the city of Bushehr Iran.

Khadijeh Arjmandifar received the B.S. in Electrical Engineering from Islamic Azad University Bushehr Branch in 1999 in Iran and M.S. degrees in Computer Engineering from Islamic Azad University Dubai Branch in 2016, respectively **.**Since 1999, she has been managing the IT department of management of the social security organization in the city of Bushehr Iran.