Journal of Engineering and Applied Sciences 15 (5): 1127-1132, 2020
ISSN: 1816-949X

Online Intrusion Detection System using C4.5 Algorithm with Honeypot

Ismael Salih Aref, Ziyad Tariq Mustafa, Iraq Ali Hussain and Samah Jalil Sabaa
Department of Computer Science, College of Science, University of Diyala, Baqubah, Iraq

Abstract: The computer security system is a protection for computers which is similar to the immune system in the human body. It includes the protection of all operations and resources within the computer and it prevention the abuse of intruders who aim to tamper with the security of the computer system. There are many important methods to protect computer systems such as intrusion detection system and honeypot system. The intrusion detection system is divided into abuse detection and anomaly detection. Detection of anomalies is an important method to detect new or unknown attacks. In this research, a robust and integrated system based on the detection system depending on the famous classification algorithm (C4.5) and the honeypot system. The proposed system is used to lure attackers and keep them away from the product system. Intrusion detection system is built and then several tests are conducted to show the detection capabilities that could be achieved. The experiment result of proposed system explain which services are most vulnerable to attack and what weaknesses are present in the product system.

Key words: Honeypot, intrusion detection system, decision tree, security, robust, computer

INTRODUCTION

The advancement of internet technology and computers helps in increasing the spread of internet services to many different places where the use of computers and internet service became available in home or business. Because of this continuous development, security loopholes have increased significantly, necessitating the development of security methods to prevent attacks that usually target vulnerable systems. The objective of security includes protection of computers, networks, software, information and property from theft, corruption or alteration, while allowing the information and property to remain accessible and productive to its intended users. Computer security is the protection afforded to an automated information system in order to attain the applicable objectives of preserving the integrity, availability and confidentiality of information system resources (includes hardware, software, firmware, telecommunications, information and data). Network security consists of many policies dedicated by the network administrator to detect and prevent modification, misuse or access to network resources (Herrmann, 2001; Stallings, 2003).

Many security methods are used to provide protection such as intrusion detection, honeypot trap system and firewall. All these share the same goal of protecting and maintaining the services and information provided to users. Intrusion detection systems monitor and analyze data passing through the network. In case of abnormal data traffic, these systems launch a warning to the network administrator which in turnper forms the necessary action by preventing unauthorized access and restricting the movement of data or other defensive means (Rout and Mohanty, 2015). Honeypot systems are a technology used to trick attackers to attack them, thus, record all the events and actions and then store them for analysis. The main objective of honeypot is to know the new methods and behavior of the attacker to take advantage of this information in the manufacture of a database for protection and defense systems.

MATERIALS AND METHODS

Background

Intrusion detection system: An intrusion is defining as "any collection of actions that attempt to compromise the availability, integrity or confidentiality of a resource". An intrusion detection system is a set of mechanisms and associated techniques that aim to monitor network or computer-host activity in order to detect and react to any attempted attack or intrusion (Ghosh *et al*., 2015; Cheswick, 1992). There have been many IDSs advanced to discover network attacks. Performance of IDS has become a concern as the researchers continue searching for an intrusion detection technology with high detection accuracy (Abadeh *et al*., 2011). IDS is an emerging area of research in computer security and network with growing usages of internet and intranet in everyday life. It can identify the user's activity as either normal or abnormal (Intrusion) and protect system for unauthorized users or attackers (Shrivas and Dewangan, 2014). IDS classify according to detection mechanism into two types, misuse detection and anomaly detection. Misuse detection work by matching known malicious pattern (signature) with actual behavior recorded in audit trail (Ghorbani *et al*., 2009). Anomaly based detection approaches initially models the normal network behavior

Corresponding Author: Ismael Salih Aref, Department of Computer Science, College of Science, University of Diyala, Baqubah, Iraq

during the training phase and then deploys the learned model to monitor the network traffic for sign of intrusions.

Decision tree: Decision trees are a very effective method of supervised learning. It aims to partition the dataset into groups as homogeneous as possible in terms of the variable to be predicted. It takes as input a set of cases or examples and create a tree data structure as output. Each case (or object) is described by a collection of feature (or attributes) which can have symbolic or numeric values. Associated with each training case (object) a label represents the class name. There are many algorithms use to contract decision tree; one of them is C4.5 algorithm. Among decision tree algorithms, C4.5 is probably the most popular in the machine learning community. The C4.5 algorithm uses a splitting criterion based on the information gain ratio. The idea is to partition the training set in such a way that the information needed to classify a given example is reduced as much as possible. The C4.5 have been used as classifier for numerous real-world domains, many of these domains, the trees produced by C4.5 are both accurate and small, resulting in reliable, fast classifiers (Hssina *et al*., 2014; Chen *et al*., 2006; Quinlan, 2014).

Honeypot: A honeypot is a closely monitored computing resource that administrator want to be probed, compromised or attacked. More precisely, a honeypot is "an information system resource whose worth lies in unauthorized or illicit use of that resource". The phrase information system resource is generally defined intentionally, so that, the honeypot can be any type of computer resource. It can be a workstation, printer, file server, mail server, router, any network device or even an entire network. It is important to denote that honeypots do not contain valuable data. Instead, they contain some type of fake data. Therefore, honeypots are the security resources that have no production value, no resource or person should be communicating with them. As such, any activity sent their way is a suspect by nature. Any traffic sent to the honeypot is most likely a scan, probe or attack (Ptacek and Newsham, 1998; Joshi and Sardana, 2011).

Honeypot can be classify based on level of interaction into low interaction and high interaction. Low interaction honeypot characterized by its limited interaction with the intruder and emulate specified services like FTP or HTTP with logging unit to record any connection to them. High interaction honeypot are based on real operating systems. The main purpose of high interaction honeypot is to provide complete access to real operating system for attacker to interaction where nothing is emulated or restricted (Joshi and Sardana, 2011; Spitzner, 2003; Mahajan *et al*., 2016).

Literature review: Recently, many researcher have been studying about intrusion detection system and honeypot. Some of many studying that are attached to the idea of this study; firstly, Ghosh *et al*. (2015) used a system for automated generation of attack signature for network intrusion detection system. This system applies pattern-recognition techniques and protocol conformance check (they examine IP, TCP, UDP headers and payload data) to the network traffic captured by honeypots. Secondly, Artail *et al*. (2006) apply Honeyd, honeynets and snort for build a hybrid honeypot approach combine both the high and the low interaction honeypots in one framework to provide more information about intruder's behavior. Thirdly, Khosravifar and Bentahar (2008) proposed a new architecture composed of distributed agents and honeypot. In this system alarming adversaries, initially detected by IDS (using snort program) will be rerouted to honeypot (Honeyd program) for more investigation. If the result of the investigation proved that the alarm caused by IDS is wrong, the connection will be forwarded to the original destination in order to continue the previous interaction. Finally, Singh and Ramajujam (2009) combine IDS and honeypot for increase the security and reliability of network. This system attempt to load balance between network performance (throughput and latency) and tools for provides security (IDS and honeypot). Load balancer receives the incoming packet, open TCP connection to IDS process and send the content of packet over that connection. IDS process check the packet and send Boolean result to load balancer.

Proposed architecture: The proposed system uses misuse intrusion detection with high interaction honeypot system. Figure 1 show the component of proposed system.

IDS architecture: The structure of intrusion detection system consists of a set of models and each model will perform a specific function. The models are complementary to each other, so that, the outputs of the first model will be used in the second model, the third model will use the output of the second unit and so on. Until final decision is created on the nature of the data that passed from network to a system to be protected. Figure 2 illustrates the proposed intrusion detection system with its main constituent models.

Sensor and monitor: This part is responsible for capturing the data stream from the Network Interface Card (NIC) and accumulating them in buffer with determined sizes (two buffers, one for save and the second for hold) and delivering the content of buffer to packet decoder which in turn will determine the header and payload of packets.

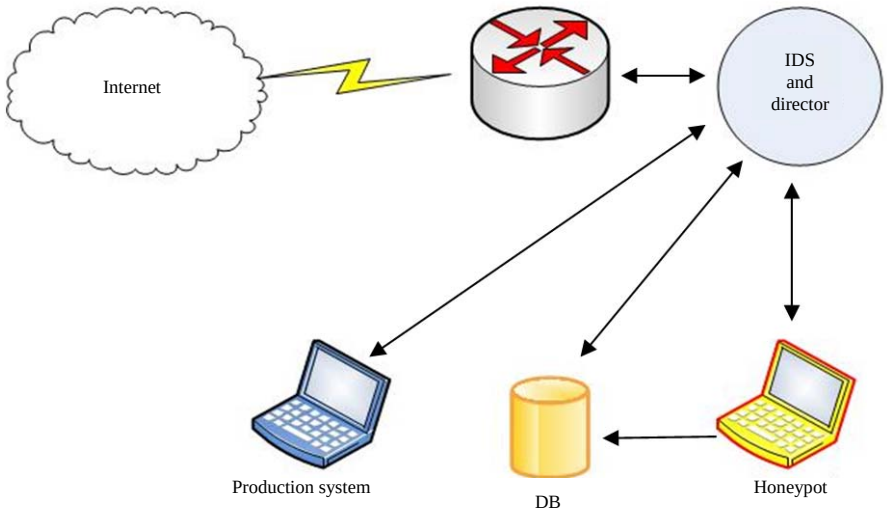

Fig. 1: Proposed IDS with honeypot block diagram

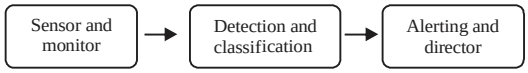

Fig. 2: Proposed IDS architecture block diagram

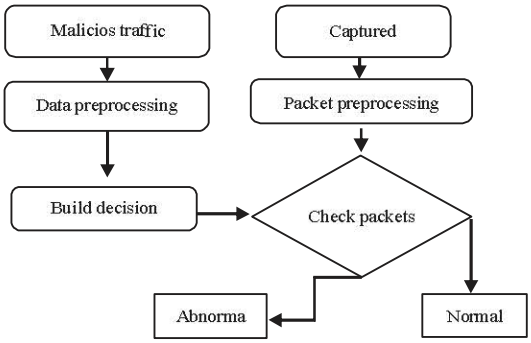

Fig. 3: The detection and classification

Detection and classification: This part detects and classifies network traffic into normal or abnormal. Detection process depend on the classification approach that uses the C4.5 algorithm for create decision tree and extract rule from it to create final decision. Figure 3 illustrate the steps of detection and classification parts.

Alerting and Director: The detection part makes the final decision on the nature of the packets which passing to computer and the output of the decision either to be normal or abnormal result and here comes the role of alert and director model. The alarm unit create alert notification to administrator about the nature of the data passing for taking the necessary actions. In addition, it stores the IP address of suspicious packets in a dedicated database. A director is a program that directs the traffic of data coming from the internet either to the product system or to the honeypot system based on IP address extracted from data traffic (normal or abnormal traffic). The director has a list which contain a suspicious IP addresses and updates its list content periodically based on the table in the database (from alert unit) used to store the IP. Figure 4 shows the role of director in IDS.

Honeypot system architecture: The honeypot is one of the main component of the proposed system. The main function of the honeypot is to lure attacker and protect the product system by deceiving the attackers with a false system interacting with them, so, it gives more time for administrator to study the nature and extent of the impact of the attack. In addition, it used to collect information and more details about the tools and methods used by attackers in the penetration process. By a trap, the administrator discover which services are most vulnerable to attack and what loophole does the attacker exploit.

The honeypot can be classifying into high and low interaction as mention previously, high interaction honeypot is deployed in real host with real operating system and real service to provide more flexibility environment for attract many attackers. In general, the honeypot consists of the following services:

- HTTP server
- FTP server
- DNS server
- Telnet server

It also contains analysis and log module used for analyze event, generate logs file and save them in database. Figure 5 illustrates the general structure of the honeypot trap in the proposed system. The information

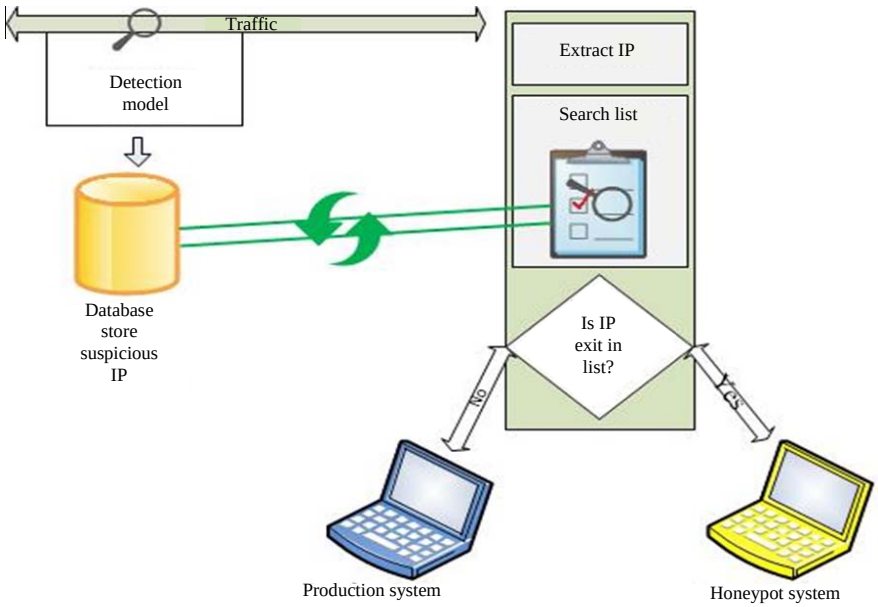

Fig. 4: The role of director

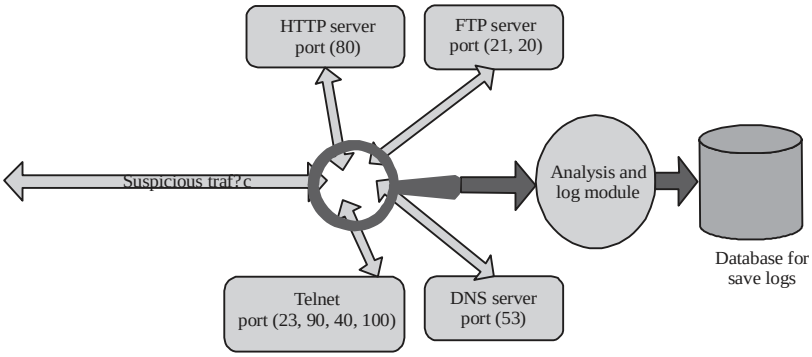

Fig. 5: The general structure of honeypot

obtained from the analysis and log module is used to detect the geographical location for IP addresses of computers devices which tried to connect with honeypot system.

RESULTS AND DISCUSSION

In this section, the experimental results of the proposed system are presented. Firstly, the honeypot installed and configured, after that run all the services mentioned early and activate the monitoring and analysis unit to record all events and take advantage of the information in creating a new database used for the detection module. Once the information is obtained, it will be analyzed and reviewed to present the most vulnerable services, as well as addresses of the computers that launching an attack.

The detection module or classifier is first constructed and then checked to verify that the unit is properly functioning. The database (NSL-KDD) used in the construction and testing of the classifier. The database is divided into two parts, the first is training data that will be used to construct the classifier and the other part will be used for test the classifier to show the accuracy of the results obtained from it. Testing data differs from training data and this difference will show the ability of the classifier to distinguish between new and unknown patterns. Testing on training data will reveal the classifier's ability to distinguish between known data. Therefore, the testing process will be done on the training and testing data.

The work will consider different numbers of data (90 and 80% for training and 10 and 20% for testing) and different number of features (all features and features selected by gain ratio). There are three measurements used to compute efficiency of the classifier to distinguish between normal and abnormal traffic data. These measurements are Accuracy (A) Detection Rate (DR) and False Positive (FP) with Eq 1, 2 and 3), respectively:

Table 1: Number of connection to honeypot from specific country

Country	No. of IPs
United States	651
Canada	17
China	10
Mauritius	6
United Kingdom	6
Spain	4
Germany	3
France	2

Table 2: Ports with highest number of connections

Connections	Local honeypot port
72	80
152	1422
168	43
254	82
471	1375
920	78
1208	62
5587	54
16416	42
34968	1442

$$A=\frac{\sum \text{true classify}}{\text{number of record}}\times 100\% \quad (1)$$

$$DR=\frac{\sum \text{true classify as abnormal}}{\text{number of record}}\times 100\% \quad (2)$$

$$FP=\frac{\sum \text{missclassify}}{\text{number of record}}\times 100\% \quad (3)$$

Test result for honeypot system: After collecting and analyzing information, a statistic calculation is to show the number of computers that have entered the honeypot, the number of services used and which protocols are most used. Thus, the statistical results will be divided into two parts:

- IP address and locations statistics
- Services usage statistics

IP address and locations statistics: In this section, the IP addresses of the computers that have accessed the honeypot are reviewed as well as the locations of those addresses. Knowing the sites that have reached the honeypot will give administrator an idea to increase security on these locations. Table 1 and Fig. 6 show how many IP addresses of devices from specific country connected with honeypot.

Services usage statistics: Knowledge of services most vulnerable to attack helps to increase attention to these services and provide maximum protection. Therefore, this statistic was conducted on services that most used in honeypot. Table 2 shows the statistics of the port of services that are used in the honeypot.

Table 3: The results of test that are applied on detection unit in test dataset

No. of data	Feature number	False positive rate (%)	Detection rate (%)	Accuracy (%)
5000	34	0.0051	99.82	99.58
7400	34	0.0051	99.799	99.56
9999	34	0.008	99.800	99.62
12498	34	0.0049	99.775	99.54
24996	34	0.0053	90.803	99.55

Table 4: Show the results of test apply detection unit on train dataset

No. of data	Feature number	False positive rate (%)	Detection rate (%)	Accuracy (%)
24995	34	0.0026	99.86	99.748
37493	34	0.0019	99.861	99.773
39992	34	0.002	99.857	99.765
42492	34	0.0021	99.882	99.783
44991	34	0.0019	99.878	99.789

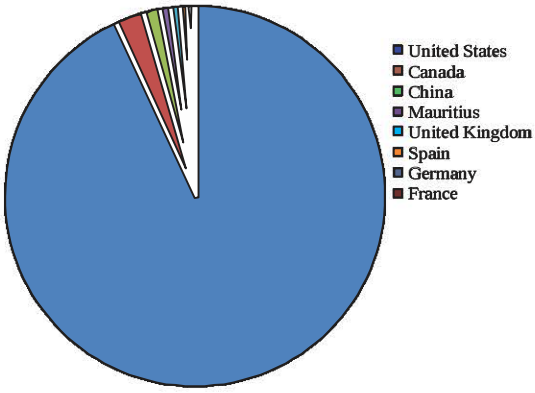

Fig. 6: The chart of distribution of visits on the honeypot

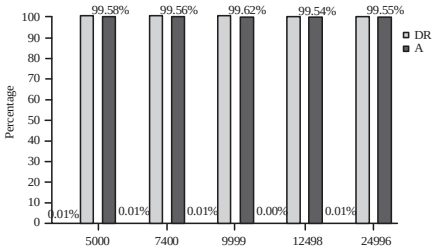

Fig. 7: Results of detection rate, false positive rate and accuracy (test data)

Test result for detection module: Many tests have been applied on the detection unit. This will help in revealing the ability of the detection unit to distinguish between different types. Table 3 and 4 explain the results of test that are applied on detection unit intest dataset and train dataset. Figure 7 and 8 display results of detection rate, false positive rate and accuracy for test and train dataset.

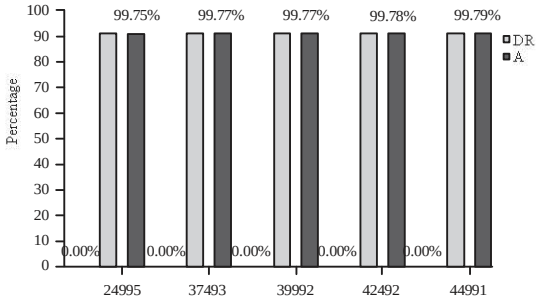

Fig. 8: Results of detection rate, false positive rate and accuracy (Train data)

CONCLUSION

Some points can be inferred from the proposed researcher. Using the C4.5 classifiers for IDS architecture allows the system to adopt new environments. This makes the proposed system able to detect unknown attack anomalies. The training of the C4.5 classifier requires a very large amount of data to ensure that the results are accurate. Also, there was some kind of compromise between increase of classification levels and the percentage of detection. Thus, a tradeoff is required.

Using many services in honeypot can attract many attackers to compromise it. And the usage of IDS with Honeypot system can provide not only more security but also more information about attackers in addition to unleashed weakness for production system.

REFERENCES

Abadeh, M.S., H. Mohamadi and J. Habibi, 2011. Design and analysis of genetic fuzzy systems for intrusion detection in computer networks. Expert Syst. Appl., 38: 7067-7075.

Artail, H., H. Safa, M. Sraj, I. Kuwatly and Z. Al-Masri, 2006. A hybrid honeypot framework for improving intrusion detection systems in protecting organizational networks. Comput. Secur., 25: 274-288.

Chen, Y., Y. Li, X.Q. Cheng and L. Guo, 2006. Building efficient intrusion detection model based on principal component analysis and C4. 5. Proceedings of the 2006 International Conference on Communication Technology, November 27-30, 2006, IEEE, Guilin, China, pp: 1-4.

Cheswick, B., 1992. An evening with berferd in which a cracker is lured, endured and studied. Proceedings of the Conference on Winter USENIX, January 20-24, 1992, San Francisco, California, pp: 20-24.

Ghorbani, A.A., W. Lu and M. Tavallaee, 2009. Network Intrusion Detection and Prevention: Concepts and Techniques. Vol. 47, Springer, Berlin, Germany, ISBN:978-0-387-88770-8, Pages: 211.

Ghosh, P., A.K. Mandal and R. Kumar, 2015. An Efficient Cloud Network Intrusion Detection System. In: Information Systems Design and Intelligent Applications, Mandal, J.K., S.C. Satapathy, M.K. Sanyal, P.P. Sarkar and A. Mukhopadhyay (Eds.). Springer, New Delhi, India, ISBN:978-81-322-2249-1, pp: 91-99.

Herrmann, D.S., 2001. A Practical Guide to Security Engineering and Information Assurance. 1st Edn., Auerbach Publications, Boca Raton, Florida, USA., ISBN:9781420031492, Pages: 408.

Hssina, B., A. Merbou, H. Ezzikouri, M. Erritali, 2014. A comparative study of decision tree ID3 and C4.5. Int. J. Adv. Comput. Sci. Applic., 2104: 13-19.

Joshi, R.C. and A. Sardana, 2011. Honeypots: A New Paradigm to Information Security. CRC Press, Boca Raton, Florida, USA., ISBN-13:978-1-4398-6999-4, Pages: 325.

Khosravifar, B. and J. Bentahar, 2008. An experience improving intrusion detection systems false alarm ratio by using honeypot. Proceedings of the 22nd International Conference on Advanced Information Networking and Applications (aina 2008), March 25-28, 2008, IEEE, Okinawa, Japan, ISBN:978-0-7695-3095-6, pp: 997-1004.

Mahajan, S., A.M. Adagale and C. Sahare, 2016. Intrusion detection system using raspberry pi honeypot in network security. Intl. J. Eng. Sci., 6: 2792-2795.

Ptacek, T.H. and T.N. Newsham, 1998. Insertion, evasion and denial of service: Eluding network intrusion detection. Technical Report T2R-0Y6, Secure Networks, Calgary, AB, Canada.

Quinlan, J.R., 2014. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, Burlington, Massachusetts, USA., Pages: 299.

Rout, G.P. and S.N. Mohanty, 2015. A hybrid approach for network intrusion detection. Proceedings of the 2015 5th International Conference on Communication Systems and Network Technologies, April 4-6, 2015, IEEE, Gwalior, India, ISBN:978-1-4799-1797-6, pp: 614-617.

Shrivas, A.K. and A.K. Dewangan, 2014. An ensemble model for classification of attacks with feature selection based on KDD99 and NSL-KDD data set. Intl. J. Comput. Appl., 99: 8-13.

Singh, R.K. and P. Ramajujam, 2009. Intrusion detection system using advanced honeypots. Intl. J. Comput. Sci. Inf. Secur., 2: 1-9.

Spitzner, L., 2003. Honeypots: Tracking Hackers. 2nd Edn., Addison-Wesley Company, Boston, Massachusetts, USA., ISBN: 9780321108951, Pages: 452.

Stallings, W., 2003. Cryptography and Network Security Principles and Practice. 3rd Edn., Prentice-Hall of India Pvt. Ltd., India.