

An Intelligent Cyber Threat Classification System

Man-Ling Liao, Chiao-Lin Yu, Yi-Cheng Lai, Shih-Ping Chiu and Jiann-Liang Chen

Department of Electrical Engineering, National Taiwan University of Science and Technology

Taipei, Taiwan

m11007502@mail.ntust.edu.tw, m10907505@mail.ntust.edu.tw, m11007505@mail.ntust.edu.tw, lchen1215@gmail.com

Abstract— This study uses TF-IDF (Term Frequency-Inverse Document Frequency) word weighting technique for information retrieval and text mining. Collect information about attackers' intrusion into the system through Honeypot, and combine data classification, manual labeling, machine learning, and K-means algorithm to provide reliable and labeled data sets with labels provided by third-party frameworks. The proposed method is mainly divided into three parts: A. Using the K-means algorithm to classify the raw data collected by Honeypots into different clusters. B. Manually label each cluster to provide reliable information according to MITRE ATT&CK. C. Each keyword score in the payload is given using TF-IDF technique and sorted into the session table. In this study, 68,936,206 packets of data are used as test data, and the proposed method can cluster the data with an accuracy rate of 99.5%.

Keywords — TF-IDF Technology, Honeypot, K-means Algorithm, MITRE ATT&CK, Cyber Threat

I. INTRODUCTION

The Internet has played an enormous role in our everyday lives. We use the Internet intensively for entertainment, online shopping, banking, education, social networking, etc. Many attackers engage in Internet crimes and computer communication attacks, for example, Phishing websites. The attacks can affect companies and organizations, leaking their information and disrupting the operating system. Once the system has been attacked, the attacker can perform a more severe internal attack.

In addition to using detection tools to detect suspicious behavior of attackers in today's technology, many recent studies also incorporate artificial intelligence algorithms to identify attacker behavior [1]-[2]. The proposed method was designed to classify issue-related actions based on AI technology and the Intelligent Threat Classification System.

This study uses the machine learning algorithm to deal with the group processing of malicious attacks in the field of information security, effectively and quickly sending the results of the grouping to the identification equipment of the information security to identify the attack behavior.

The study uses TF-IDF technology, a common word weighting technique used in information retrieval and word mining, to collect suspicious behavior instructions through Honeypots and evaluate the importance of a single word to all instructions. The volume of a comment increases with it in a packet. Increases proportionally to the number of times it occurs, but at the same time decreases inversely proportional to the number of times it appears in all instructions. Use this to understand what types of attacks these suspicious behaviors are in individual packets [3]-[4].

This study also uses the unsupervised learning K-means algorithm to cluster the suspicious behavior instructions collected by Honeypots. The method is to calculate the distance between each data as the basis for clustering and to determine each cluster in which cluster by repeatedly calculating and changing the center point of each cluster. After clustering, each cluster will be marked according to the MITRE ATT&CK framework to identify the attack behavior it belongs to so as to improve the credibility of the data. The next chapter will introduce each in detail.

The results obtained by the K-means algorithm are weighted for malicious instructions through TF-IDF text weighting technology so that the results can present the most aggressive part of complex instruction. The techniques mentioned above will be described in detail in the next section.

II. RELATED WORKS

This study uses Cowrie Honeypot, a moderately to highly interactive SSH, and Telnet Honeypot is used to record the logging commands left by the attacker in the Shell. The Honeypots simulate a UNIX system in Python as a proxy for SSH and Telnet and then observe what an attacker would do to another system [5]. Disguised as a valuable computer device, Cowrie Honeypot can effectively observe all the behaviors of attackers on Honeypots. Bryson et al. showed that Cowrie Honeypot is an effective system for collecting samples [6]. S. Dowling et al. [7] utilize the dataset collected by Honeypots to determine attack types and corresponding temporal patterns. It proposed a Honeypot that can modify its security level to increase attack vectors at different times of the day.

MITRE ATT&CK is an information database platform that records information security attacks and information sharing

[8]. Currently, most of the attack methods on the Internet can be found in its tags. Since MITRE ATT&CK will update the intelligence data at regular intervals, it is more reliable to handle the classification and labeling of this study. R. Kwon et al. proposed a threat dictionary based on the MITRE ATT&CK and NIST cybersecurity framework [9].

TF-IDF is a commonly used word weighting technique for information retrieval and word mining. It was used to assess the importance of a word to a file. This study adopts this technique to evaluate the more dangerous keywords in the payload and present the more dangerous commands in the whole Session. Also, search engines often use this technology to measure the relevance between files and user queries.

K-means algorithm is a cluster analysis method often used in data mining [10]. There are many applications of K-means, such as air pollution research, student achievement analysis, etc. There is some research about K-means. S.D. Roy et al. [11] proposed a decentralized Intrusion Detection System that combines K-means and identification sensors to measure and control the degree of data anomaly in the signal. Y. Li et al. [12] proposed an ant K-means (AK) algorithm based on K-means, which uses probability to locate objects in a cluster. The likelihood is updated by pheromone. This study uses the K-means algorithm to group the payload. Since the difference in each cluster is small, it is easier to sort out the relevant information in the data. It can be more efficient to understand the behavior pattern of the attackers after sorting it into a Session Table.

III. PROPOSED SYSTEM

This study proposes a system to use Honeypots to collect suspicious attack behavior packets as a dataset for data processing to present reliable and meaningful informational tagging of aggression to the public. The flow includes three steps: A. After collecting the data through Honeypots, integrating the K-means algorithm to cluster the simulated payload in each cluster. B. Manually tag each cluster's attributes by MITRE ATT&CK. C. Calculate the TF-IDF value of the individual payload and save the labeled data, data-related information, and TF-IDF calculated scores into the Session Table. The proposed system flow is shown in Figure 1.

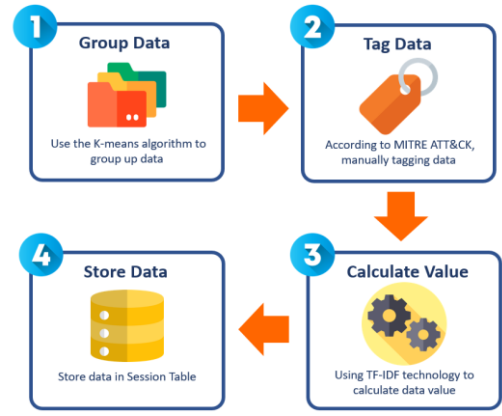


Figure 1. Proposed System Flow

A. After Honeypots receive the suspicious packets accessed by the attacker, this system will use the K-means algorithm to convert all payloads into values in the vector space individually. K's value is the number of clusters. After setting parameter K, sprinkle K center points in the entire vector space. These payloads are clustered according to the distance close to the center point.

K-means algorithm is an unsupervised learning algorithm, as shown in Figure 2. The processing flow includes the following:

1. The parameter K is first set, used as the number of cluster centers in the vector.
2. The cluster center is scattered in the vector space for each data. Calculate the shortest Euclidean distance between each data to all cluster centers.
3. Take the cluster with the shortest distance for each data and classify it as that.
4. Finally, repeat steps 2 and 3. Finish until all data points stay the same from the cluster center.

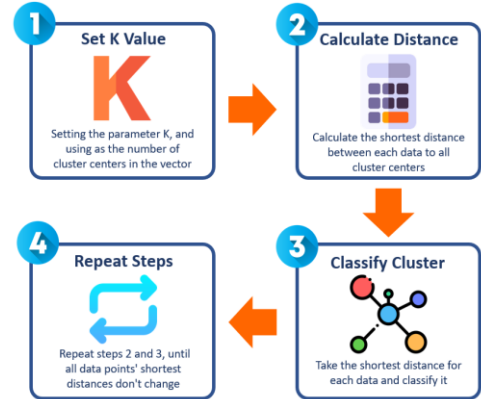


Figure 2. K-means Algorithm

B. After clustering, perform labeling actions of Tactic, Technique, and Subtechnique in each cluster according to the MITRE ATT&CK framework. Tactic represents the main action of an event and describes the broad categories of attacks that attackers will use, shown in Table 1, which lists all the current MITRE ATT&CK Tactic items and their descriptions.

Technique represents different event categories in Tactic, as shown in Table 2, which describes the names and descriptions of some Techniques in Tactic Execution.

While Subtechnique represents the sub-items in action, take Technique Command and Scripting Interpreter (T1059) as an example; as shown in Table 3, Subtechnique PowerShell (T1059.001) indicates that attackers may abuse PowerShell commands and scripts to execute. As shown in Table 4, by labeling the attack behavior through MITRE ATT&CK, one can know which types of attack behavior are in each cluster. The process and purpose of the method reflect the threat level of each attack method, and finally, each group is marked with a Malicious score to show the degree of harm. Regarding the standard of Malicious score, personnel from three groups of different institutions in Taiwan have made a unified standard for it. Each score criterion is shown in Table 5.

TABLE 1. The current MITRE ATT&CK Tactic

Tactic	Description
Initial Access TA0001	Attackers use various network access technology on victim devices.
Execution TA0002	Attackers use the execution command on the victim's device.
Persistence TA0003	Attackers maintain access to victim devices, remotely reboot, change credentials, and more.
Privilege Escalation TA0004	Attackers try to gain elevated privileges on the victim's device.
Defense Evasion TA0005	Attackers try to hide their tracks on a victim's device.
Credential Access TA0006	Attackers steal credentials such as account passwords on victim devices.
Resource Development TA0042	Attackers achieve their goals by creating, buying, destroying, stealing, and more.
Reconnaissance TA0043	Attackers actively or passively collect victim device information to achieve their goals.

TABLE 2. Some Techniques in Tactic Execution

Tactic: Execution (TA0002)	
Technique	Description
Command and Scripting Interpreter T1059	Attackers may use automatic or manual methods to execute commands.
Container Administration Command T1609	Attackers may use container management services to execute instructions.
Deploy Container T1610	Attackers may deploy a container environment to the victim device to execute a command.
Exploitation for Client Execution T1203	Attackers may exploit client application vulnerabilities in victim devices to execute commands.
Inter-Process Communication T1559	Attackers may exploit inter-process communication (IPC) to execute commands.

TABLE 3. The detail in Technique: Command and Scripting Interpreter (T1059)

Technique: Command and Scripting Interpreter (T1059)		
Subtechnique	Name	Description
T1059.001	PowerShell	Attackers may use PowerShell commands to execute.
T1059.002	AppleScript	Attackers may use AppleScript to execute.
T1059.003	Windows Command Shell	Attackers may use the Windows command shell to execute.
T1059.004	Unix Shell	Attackers may use Unix shell commands and scripts to execute.
T1059.005	Visual Basic	Attackers may use Visual Basic (VB) to execute.
T1059.006	Python	Attackers may use Python commands and scripts to execute.
T1059.007	JavaScript	Attackers may use various implementations of JavaScript to execute.

T1059.008	Network Device CLI	Attackers may use scripting or built-in command line interpreters (CLI) on network devices to execute malicious commands and payloads.
-----------	--------------------	--

TABLE 4. Used MITRE ATT&CK Framework

Cluster ID	Tactic	Subtechnique	Malicious Score
1	TA0002	T1059.004	0.8
2	TA0002	T1059.004	0.8
3	TA0002	T1059.004	0.8
4	TA0002	T1059.004	0.8
5	TA0007	T1120.000	0.6
6	TA0002	T1059.004	0.8
7	TA0002	T1059.004	0.8
8	TA0002	T1059.004	0.8
9	TA0001 TA0002	T1190.000 T1059.004	0.8
10	TA0002	T1059.004	1
11	TA0007	T1087.001	0.6
12	TA0002	T1059.004	0.8
13	TA0001 TA0002 TA0007 TA0040	T1190.000 T1059.004 T1087.001 T1485.000	0.8
14	TA0001 TA0002 TA0040	T1190.000 T1059.004 T1485.000	0.6
15	TA0007 TA0002	T1087.001 T1059.004	1
16	TA0002	T1059.004	0.8
17	TA0002	T1059.004	0.8
18	TA0002	T1059.004	1
19	TA0002	T1059.004	0.8
20	TA0002	T1059.004	0.8
21	TA0002	T1059.004	0.8
22	TA0002	T1059.004	0.8
23	TA0002	T1059.004	0.8
24	TA0002	T1059.004	0.8
25	TA0002	T1059.004	0.8
26	TA0002	T1059.004	0.8
27	TA0002	T1059.004	0.8
28	TA0040 TA0002	T1485.000 T1059.004	0.8
29	TA0001	T1190	0.2
30	TA0006	T1110.000	0.6
31	TA0002	T1190.000 T1059.004	0.6

TABLE 5. The standard of Malicious score

The standard of Malicious score	
0	The command is a blank, line feed, and other meaningless commands.

0.2 – 0.4	The command tries to connect.
0.6	Brute force crack account password.
0.8 - 1	Enter commands, Exploit, and other actions on the system.

C. TF-IDF technology, shown in Figure 3, for word processing.

1. Term Frequency (TF) calculates the number of times a word appears in a single Payload.
2. Inverse Document Frequency (IDF) estimates the number of times the word appears in all Payloads.
3. Finally, the two are multiplied to obtain a weighted score.

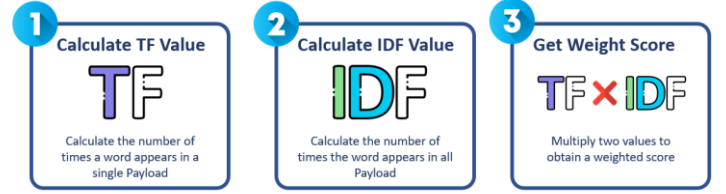


Figure 3. TF-IDF Technology

The mathematical formula for TF-IDF is as follows:

$$TF_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \quad (1)$$

n_{ij} is the number of occurrences of the word in file d_j , and the denominator is the sum of the occurrences of all words in file d_j .

$$IDF_i = \lg \frac{|D|}{|\{j : t_i \in d_j\}|} \quad (2)$$

$|D|$ is the total number of files in the database, and the denominator represents the number of files containing the word t_i .

$$TFIDF_{ij} = TF_{ij} \times IDF_i \quad (3)$$

Finally, multiply the two to obtain the TF-IDF weight value.

IV. PERFORMANCE ANALYSIS

In this study, the JSON files provided by HoneyPot of 9 ISPs in Taiwan were converted into features that can be provided to the K-means algorithm after data preprocessing. After using the K-means clustering algorithm and labeling features by MITRE ATT&CK, Table 6 is the proportion ranking table collected by Cowrie HoneyPot for a week. It can be observed from this table that in the clusters by the K-means algorithm, the same Cluster ID represents a similar or the same payload. After checking every payload under Cluster ID, Cluster ID: 31 and Cluster ID: 30 represent “Brute force account password” and “Simple SSH connection.” The command of “connection” occupies a large part of the attack behavior, but the attack behavior that accounts

for a small proportion has a higher Malicious score. The Malicious score is inversely proportional to the ratio of attack behavior. If a unique analysis is required for high-threat attack behavior in the future, the tail data with a small proportion will be the group that must be paid attention to.

TABLE 6. Week-long SSH Interaction Data

Cluster ID	Packet number	Percentage	Malicious Score
31	51648542	74.90%	0.6
30	17095046	24.80%	0.6
5	122275	0.18%	0.6
11	38150	0.06%	0.6
18	8313	0.01%	1
13	6917	0.01%	0.8
24	5202	0.01%	0.8
22	3675	0.01%	0.8
3	2892	0.004%	0.8
14	2253	0.003%	0.6
12	652	0.001%	0.8
1	549	0.001%	0.8
15	343	0.0005%	1
16	307	0.0004%	0.8
17	305	0.0004%	0.8
4	187	0.0003%	0.8
9	148	0.0002%	0.8
2	119	0.0002%	0.8
7	117	0.0002%	0.8
10	57	0.0000083%	1
20	43	0.0000062%	0.8
6	42	0.0000061%	0.8
19	28	0.0000041%	0.8
8	13	0.0000019%	0.8
28	11	0.0000016%	0.8
23	10	0.0000015%	0.8
21	7	0.000001%	0.8
26	2	0.0000003%	0.8
25	1	0.00000015%	0.8

Table 7 shows the DataFrame format of the SSH Session Table. It can be observed that this table contains each Session's Session ID, the results of MITRE ATT&CK labeling, the risk level of the Malicious score, and the scores calculated by the TF-IDF weighting technique.

TABLE 7. SSH Session Table for 3 Sampled Data

ID	Tactic	Technique	Sub Technique	Malicious Score	Packet Score
1	TA0006	T1110	.001	0.6	0.888
	TA0006	T1110	.001		
2	TA0002	T1059	.004	0.8	0.710
3	TA0001	T1190	.000	0.8	0.782
	TA0002	T1059	.004		0.646
	TA0007	T1087	.001		0.450
	TA0040	T1485	.000		0.130

This study uses the Within-cluster sum of squares (WCSS) [13] to find the appropriate K in the K-means algorithm.

The mathematical formula for WCSS is as follows:

$$WCSS = \sum_{k=1}^K \sum_{i \in A_k} ||y_i - c_k||^2 \quad (1)$$

Among them, A is the K cluster division represented by the vector y_i in the feature space of multiple dimensions, which is composed of non-null and non-overlapping clusters A_k , and each cluster has a center point c_k ($k = 1, 2, \dots, K$).

It can be seen from Figure 4. that the K value after about 20, the slope obtained by WCSS gradually slows down. After our analysis and observation of the 450,000 data, the first few large groups account for a high proportion of the total 450,000 data, and the similarity is very high. Suppose the value of K is too large. In that case, the remaining groups with small proportions will only be divided into more pieces, which is not conducive to the analysis and research of aggressive behavior. Therefore, The K value is around 30 for the K-means algorithm clustering in this study.

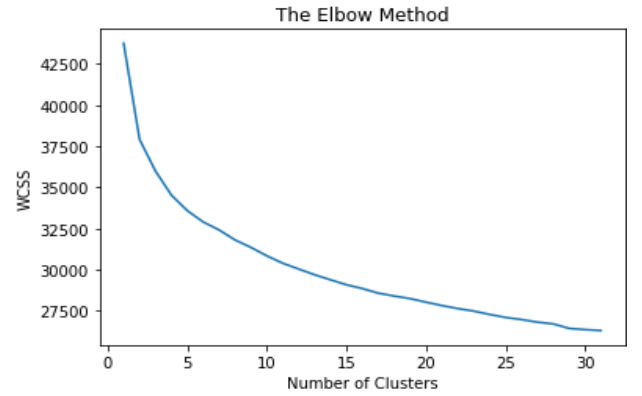


Figure 4. The Elbow Method

Table 8 is the confusion matrix of the data through the K-means algorithm. TP is the number of packets that are grouped correctly after the algorithm, FP is the number of packets that are incorrectly grouped, FN is the number of packets that can not be grouped but can be classified into clusters after the algorithm judges, and TN is the number of packets that can not be grouped after the algorithm judges the number of packets.

TABLE 8. Confusion matrix of the data through the K-means algorithm

	True Yes	True No
Predict Yes	TP 68,575,748	FP 245,758
Predict No	FN 98,942	TN 15,776

Table 9 shows the accuracy of clustering payloads using the K-means algorithm, which is as high as 99%. Very few payloads

clustered to the wrong cluster, indicating that the K-means algorithm's clustering can effectively distinguish different attack patterns.

TABLE 9. Week-long SSH Testing Information

K-means Clustering Information (SSH)	
Cluster Number	31
Packet Number	68,936,206
Abnormal Clustering	245,758
Accuracy	0.995

V. CONCLUSIONS

This study organizes and analyzes the attack information collected by Honeypots. The payloads from Honeypot are clustered through the K-means algorithm, letting similar or the same payload together. The results of the clusters are combined with MITRE ATT&CK to increase the reliability of the dataset in the study. Mark Malicious score indicates its degree of harm and calculates its TF-IDF value to know which word in this payload is hazardous. Finally, save this information in Session Table as a reliable dataset. Although the experimental results are good, there are limitations. The data used in this study are only SSH data collected by Honeypot. Although the proposed system clustered the data with an accuracy of 99.5%, it turns out that the material is labeled to provide a reliable dataset with realistic examples. Still, the experiments in this study do not apply to other occasions.

REFERENCES

- [1] A.J. Park, R.N. Quadari and H.H. Tsang, "Phishing Website Detection Framework through web scraping and Data Mining," *Proceedings of the IEEE Annual Information Technology, Electronics and Mobile Communication Conference*, pp.680-684, 2017.
- [2] B.X. Wang, J.L. Chen, and C.L. Yu, "An AI-powered Network Threat Detection System," *IEEE Access*, Vol.10, pp.54029 – 54037, 2022.
- [3] G.A. Dalaorao, A.M. Sison and R.P. Medina, "Integrating Collocation as TF-IDF Enhancement to Improve Classification Accuracy," *Proceedings of the IEEE 13th International Conference on Telecommunication Systems, Services, and Applications*, Indonesia, October 2019.
- [4] A.S. Alammery, "Arabic Questions Classification Using Modified TF-IDF," *IEEE Access*, Vol.9, pp. 95109 – 95122, 2021.
- [5] Profit from Tech, *cowrie/cowrie: Cowrie SSH/Telnet Honeypot*, <https://cowrie.readthedocs.io>.
- [6] B. Lingenfelter, I. Vakiliinia, and S. Sengupta, "Analyzing Variation Among IoT Botnets using Medium Interaction Honeypots," *Proceedings of the 10th Annual Computing and Communication Workshop and Conference*, pp.761-767, 2020.
- [7] S. Dowling, M. Schukat and H. Melvin, "Using Analysis of Temporal Variances within a Honeypot Dataset to Better Predict Attack Type Probability," *Proceedings of the 12th International Conference for Internet Technology and Secured Transactions*, pp.349-354, 2017.
- [8] B.E. Strom, A. Applebaum, D.P. Miller, K.C. Nickels, A.G. Pennington, and C.B. Thomas, *Mitre ATT&CK: Design and Philosophy*, MITRE PRODUCT, MP180360R1, 2018.
- [9] R. Kwon, T. Ashley, J. Castleberry, P. Mckenzie, and S.N.G. Gourisetti, "Cyber Threat Dictionary using MITRE ATT&CK Matrix and NIST Cybersecurity Framework Mapping," *Proceedings of the Resilience Week*, 2020.
- [10] C. Liu, Z. Gu, J. Wang, "A Hybrid Intrusion Detection System Based on Scalable K-Means+ Random Forest and Deep Learning," *IEEE Access*, Vol.9, pp. 75729 - 75740, 2021.
- [11] S.D. Roy, S. Debbarma, and A. Iqbal, "A Decentralized Intrusion Detection System for Security of Generation Control," *IEEE Internet of Things Journal*, 10.1109/JIOT.2022.3163502, 2022.
- [12] Y. Li, Y. Lu, and S. Li, "EZAC: Encrypted Zero-day Applications Classification using CNN and K-Means," *Proceedings of the IEEE 24th International Conference on Computer Supported Cooperative Work in Design*, China, May 2021.
- [13] Trupti m Kodinariya and Prashant Makwana, "Review on Determining of Cluster in K-means Clustering," *International Journal of Advance Research in Computer Science and Management Studies*, November 2013.