Sodinokibi intrusion detection based on logs clustering and random forest

Kévin Cortial*
Open Studio, F-43000, Le Puy-en-Velay, France
kcortial@openstudio.fr

Arnault Pachot
Open Studio, F-43000, Le Puy-en-Velay, France
apachot@openstudio.fr

ABSTRACT

Cyber-attacks are becoming more common and their consequences more and more disastrous. Machine learning is revolutionizing cyber security by analyzing massive amounts of data automatically. In this paper, we test the unsupervised learning method of k-means to detect the intrusion of Sodinokibi ransomware in logs. The k-means highlighted a small cluster of anomalous logs that are revealed to be the entry points of the cyberattack. This positive result allows us to consider automating of k-means, as a solution to monitor logs in real time and report abnormal behavior.

CCS CONCEPTS

• **Intrusion/anomaly detection and malware mitigation**; • **Machine learning**;

KEYWORDS

Ransomware, Cybersecurity

ACM Reference Format:
Kévin Cortial and Arnault Pachot. 2021. Sodinokibi intrusion detection based on logs clustering and random forest. In *2021 2nd International Conference on Artificial Intelligence and Information Systems (ICAIIS '21), May 28–30, 2021, Chongqing, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3469213.3469221

1 INTRODUCTION

Nowadays, cyber security is a challenging issue in the cyber space, and it has been increasing dramatically depending on computerization on different application domains including finances, industry, medical, and many other important areas. This digitalization is the consequence of a dependence on data and increases their value. So, there is a strong demand for effective intrusion detection system that is designed to interpret intrusion attempts of incoming network traffic efficiently to protect data. Intrusion detection can also be applied beyond detecting cyber- attack in noticing abnormal system behavior to identify accidents or unexpected conditions. There are two types of security systems for intrusion detection [1]:

- The detection of known and recurring cyber-attacks such as phishing. Thus, we can create training data for learning supervised algorithms. However, this type of training data is difficult to obtain in a real network environment. Moreover, as the network environment or services change, normal traffic patterns are modified. This leads to a high false positive rate.
- The anomaly detection where unknown attacks. Thus, an unsupervised algorithm can be trained with unlabeled data and can detect previously unseen attacks. This paper uses an intelligent unsupervised system to detect an unknown ransomware intrusion.

Sodinokibi is a ransomware malware with different and original intrusions in the computer network. This software encrypts personal data and then asks the owner to send money in exchange for the key to decrypt it. Sodinokibi, also known as REvil, first appeared in April 2019 and gained prominence after. The consequences of this intrusion are data theft and threats of data disclosure. Thus, the contribution of our research work in this paper are the detection of Sodinokibi ransomware. Indeed, this type of attack is quite new and is increasingly used by cybercriminals.

2 RELATED WORK

Having no clear idea about the form of Sodinokibi attacks, machine learning proved to be an effective tool to detect abnormal and unusual events. Thus, we used unsupervised models to discover underlying structures in unlabeled logs data. Each approach to implementing an intrusion detection system has its own advantages and disadvantages, a point apparent from the discussion of comparisons among the various methods. Thus, it is difficult to choose an unsupervised method to implement an intrusion detection system over the others [2].

In the scientific literature, numerous publications prove the effectiveness of machine learning methods in detecting malicious intrusions in computer systems. The k-means algorithm is the most popular in this context of cyber-attacks [3] [4]. Indeed, mine network data for anomalies based on the k-means algorithm. Thus, training data containing unlabeled logs are separated into clusters of normal and abnormal logs. Furthermore, this method could be automated to detect, in real time, intrusive activities on a computer system [5] [6]. The drawback of this existing work is the difficulty to minimize false alarm while maximizing detection and accuracy rate.

Other unsupervised machine learning methods such as Support Vector Machine (SVM) can be considered. However, the efficiency of this algorithm is proven when coupled with other methods such as hierarchical clustering [7] or complex algorithms such as Grey Wolf Optimization (GWO) [8]. These SVM methods have equivalent

*Corresponding author.

ICAIIS '21, May 28–30, 2021, Chongqing, China

ACM ISBN 978-1-4503-9020-0/21/05…$15.00
https://doi.org/10.1145/3469213.3469221

results to k-means but they are more complex to industrialize and automate. So, in this paper, we will use the k-means method.

3 CASE STUDY

An industrial group suffered a cyber-attack that allowed hackers to break into the computer network and install the Sodinokibi ransomware-type malware on numerous workstations. The affected company is trying to understand how the hackers managed to infiltrate their computer network.

The company provided us with millions of logs (event logs) from their Kaspersky antivirus (800 000 logs) and their Fortigate 300 firewall (7 000 000 logs) to analyze. In total, we had 8 files of about one million logs with about 100 variables characterizing them. Our objective was to analyze these logs post-attack to find the entry points in the system and to trace the attack. To process these huge volumes of data, we implemented a scientific approach based on machine learning.

4 METHOD

4.1 Preprocessing for logs data

The millions of logs are mostly in the form of textual and categorical variables. A preprocessing was necessary to transform and exploit these data easily. The list of preprocessing is given here:

- The replacement of missing data by a value in order to consider this information as a modality.
- The label encoding of these categorical variables into numerical data. Thus, each modality has an assigned number between 0 and $number_{classes} - 1$.
- Scaling the data to reduce the importance or underimportance of a modality due to the assignment of any number in the previous step. This scaling is done using the equation 1).

$$x\_scale = \frac{x \ - - \ min(x)}{max(x) \ - - \ min(x)} \tag{1}$$

4.2 K-means clustering algorithm

The k-means algorithm finds locally optimal solutions regarding the clustering error. This is a fast iterative algorithm that has been used in many clustering applications. This is a point-based clustering method that starts with the cluster centers initially placed at arbitrary positions and proceeds by moving the cluster centers at each step to minimize clustering error.

The main drawback of this method lies in its sensitivity to the initial positions of the cluster centers. Therefore, to obtain quasi-optimal solutions using the k-means algorithm, it is necessary to program several executions which differ in the initial positions of the cluster centers.

The k-means algorithm is a data partitioning method that divides logs into k groups often called clusters. Mathematically, k-means groups the data by trying to separate the samples into n groups of equal variances, minimizing the criterion of inertia or intra-cluster sum of squares defined by the following equation 2).

$$\sum_{i=0}^{n} \left(\min_{\mu_j \in C} \left(|x_i - \mu_j|^2 \right) \right) \tag{2}$$

The k-means algorithm divides a set of n samples x into k disjoint clusters C, each described by the mean μ_j of the cluster samples. The averages are commonly referred to as the cluster centroids.

Mainly, the experimental setup consists in choosing the number of clusters to retain. Indeed, several methods exist such as the analysis of the percentage of variance explained according to the number of clusters [9]. To achieve this, one must solve the problem of clustering k for different numbers of clusters, and then use appropriate criteria to select the most appropriate value of k. In this case, the proposed method directly provides clustering solutions for all the intermediate values of k, therefore requiring no additional computational effort. A certain number of clusters must be chosen so that adding another cluster does not result in a better modeling of the data. Specifically, when the percentage of variance explained by clusters versus the number of clusters, the firsts clusters add a lot of information (explain a lot of variance), but at some point, the marginal gain decreases. Thus, the number of clusters is chosen at this stage, when the addition of a cluster decreases little.

5 EXPERIMENTATIONS RESULTS

5.1 Cluster detection with k-means

For our case study, we applied the k-means method to computer logs. The analysis of the percentage of variance explained according to the number of clusters advises us to create two clusters of logs. An example of these results can be seen in figure 1. The axes result from non-linear combinations of the categorical variables of the logs with the t-Distributed Stochastic Neighbor Embedding method (t-SNE) [10].

Thus, the k-means algorithm can highlight groups of suspicious logs. However, we need to identify the characteristics of each cluster. This categorization allows us to label the logs in order to perform a supervised learning of the clusters. The metric used to measure the performance of the experiments is only done with the validation of the suspect logs by a cybersecurity expert. Thus, a percentage of good prediction is calculated with the conclusions of his analysis.

5.2 Cluster characterization with random forest

The decision tree forest method identifies the variables that discriminate the created clusters. Indeed, the variables that are frequently found in the nodes of the decision trees will be the discriminating variables. The Gini score measures how frequently a random element in the set would be misclassified if its label were chosen randomly, based on the distribution of labels in the subset. The Gini score is done using the equation 3) where p_i is the proportion of the samples that belongs to class k for a particular node.

$$Gini \ = 1 \ - \ \sum_{i=1}^{k} p_i^2 \tag{3}$$

To determine the set of discriminating variables, the decision tree forest algorithm performs a training on several decision trees. Thus, we can know the most discriminating variables by simulating several hundred trees. Random Forest Classifer provides the importance of the variables by calculating the number of samples that

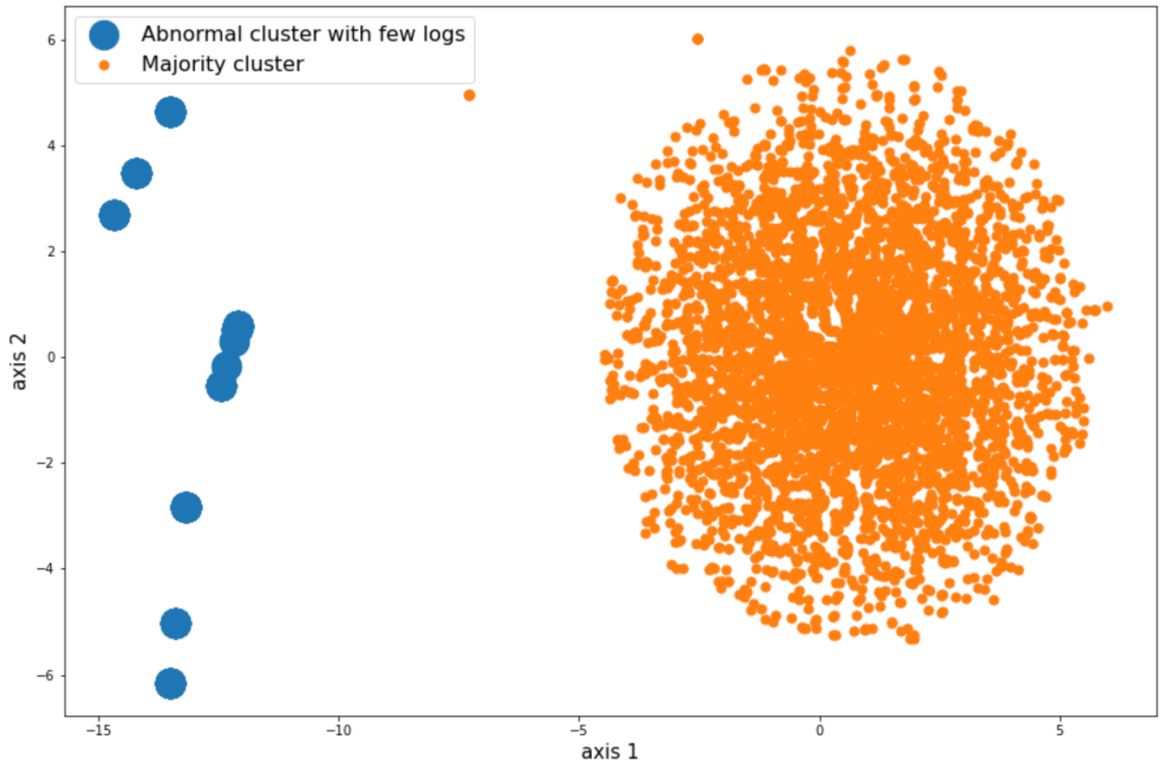

Figure 1: Example of a visualization of the clusters on the axes of t-SNE

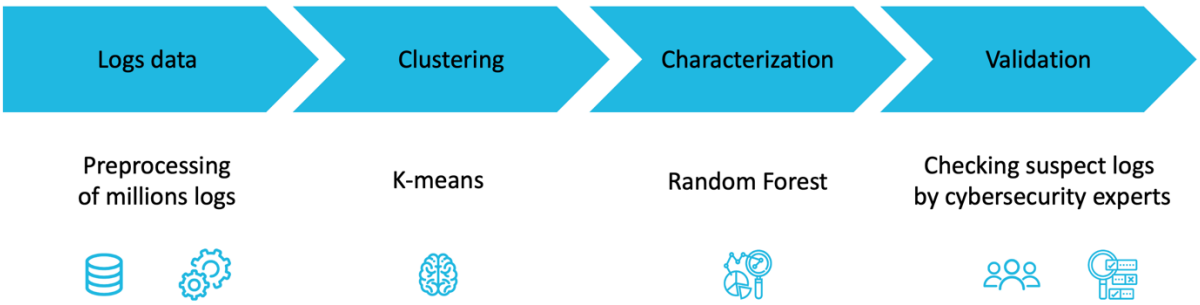

Figure 2: Workflow of the entire process

reach the node (and thus the variable in question) divided by the total number of samples. This importance value allows us to identify the most discriminating exogenous variables.

This work allows us to identify the main variables that influence the creation of clusters and therefore of atypical logs. In practice, when the number of variables is very important, this method allows to quickly identify the variables that characterize each cluster. Thus, the study of these variables allows to know the values corresponding to the clusters of suspicious logs for example.

This last step allows the completion of the workflow, figure 2, by allowing the cybersecurity experts to rule on the dangerousness or not of the highlighted logs.

6 CONCLUSION AND PERSPECTIVES

The analysis of these millions of post-attack logs by k-means allowed us to find the entry points of the ransomware into the system. We were able to provide the industry group with a list of small suspect logs. The characteristics of these logs could be determined with the random forest.

Unsupervised models, such as k-means, can automatically detect suspicious logs in millions of data sets. Thus, a machine learning system can provide a few suspicious logs that can be analyzed by a business expert. Moreover, supervised learning of these logs, now labeled, allows to better characterize these suspicious clusters using the values of some variables.

For the detection of known attacks, a second type of model is effective. Supervised algorithms can recognize whether a set of logs is aggressive or not. These models require to build a database of labeled logs. The construction of this database can be done by simulating possible attacks within a computer system or by recovering logs of proven attacks.

Automating these artificial intelligence tools would allow near real-time monitoring of events to flag abnormal behavior. Tools, such as Elasticsearch, will be able to automatically retrieve logs and apply machine learning models to detect suspicious events. Future improvements are possible such as combining the k-means method with other unsupervised methods.

In no case will an artificial intelligence method guarantee the security of systems. However, several AI-based alerts can drastically improve the effectiveness of IT security services.

ACKNOWLEDGMENTS

This paper and the research behind it would not have been possible without the help of our team at OpenStudio: Alexandre ALAIMO, Vincent CLERC, Jean-Luc MARINI, Théo PAPUT, Pierre TISSEUR and Frédéric WANG.

REFERENCES

[1] P. Gogoi, B. Borah, et D. K. Bhattacharyya, Anomaly Detection Analysis of Intrusion Data using Supervised & Unsupervised Approach, JCIT J. Converg. Inf. Technol., p. 95-110, 2010.

[2] Y. Xin *et al.*, *Machine Learning and Deep Learning Methods for Cybersecurity*, *IEEE Access*, vol. 6, p. 35365-35381, 2018, doi: 10.1109/ACCESS.2018.2836950.

[3] G. Münz, S. Li, et G. Carle, *Traffic Anomaly Detection Using K-Means Clustering*, 2007. /paper/Traffic-Anomaly-Detection-Using-K-Means-Clustering-M%C3%BCnz-Li/634e2f1a20755e7ab18e8e8094f48e140a32dacd.

[4] Z. Muda, W. Yassin, M. N. Sulaiman, et N. I. Udzir, Intrusion detection based on K-Means clustering and Naïve Bayes classification, in 2011 7th International Conference on Information Technology in Asia, juill. 2011, p. 1-6, doi: 10.1109/CITA.2011.5999520.

[5] R. Kumari, Sheetanshu, M. K. Singh, R. Jha, et N. K. Singh, Anomaly detection in network traffic using K-mean clustering, in 2016 3rd International Conference on Recent Advances in Information Technology (RAIT), mars 2016, p. 387-393, doi: 10.1109/RAIT.2016.7507933.

[6] Y. Gu, K. Li, Z. Guo, et Y. Wang, *Semi-Supervised K-Means DDoS Detection Method Using Hybrid Feature Selection Algorithm*, *IEEE Access*, vol. 7, p. 64351-64365, 2019, doi: 10.1109/ACCESS.2019.2917532.

[7] S.-J. Horng *et al.*, *A novel intrusion detection system based on hierarchical clustering and support vector machines* , *Expert Syst. Appl.*, vol. 38, n° 1, p. 306-313, janv. 2011, doi: 10.1016/j.eswa.2010.06.066.

[8] H. Lu, G. Zhang, et Y. Shen, *Cyber Security Situation Prediction Model Based on GWO-SVM*, in *Innovative Mobile and Internet Services in Ubiquitous Computing*, Cham, 2020, p. 162-171, doi: 10.1007/978-3-030-22263-5_16.

[9] G. W. Milligan et M. C. Cooper, *An examination of procedures for determining the number of clusters in a data set*, *Psychometrika*, vol. 50, n° 2, p. 159-179, juin 1985, doi: 10.1007/BF02294245.

[10] H. Zhou, F. Wang, et P. Tao, t-Distributed Stochastic Neighbor Embedding Method with the Least Information Loss for Macromolecular Simulations, J. Chem. Theory Comput., vol. 14, n° 11, p. 5499-5510, nov. 2018, doi: 10.1021/acs.jctc.8b00652.