Sunday, April 19, 2020

Machine Learning and Cryptographic Algorithms – Analysis and Design in Ransomware and Vulnerabilities Detection

Nandkumar A Niture

Information System Engineering Management

Harrisburg University of Science and Technology

Harrisburg, PA, USA

***Abstract* - The AI, deep learning and machine learning algorithms are gaining the ground in every application domain of information technology including information security. In formation security domain knows for traditional password management systems, auto-provisioning systems and user information management systems. There is another raising concern on the application and system level security with ransomware. On the existing systems cyber-attacks of Ransomware asking for ransom increasing every day. Ransomware is the class of malware where the goal is to gain the data through encryption mechanism and render back with the ransom. The ransomware attacks are mainly on the vulnerable systems which are exposed to the network with weak security measures. With the help of machine learning algorithms, the pattern of the attacks can be analyzed. Create or discuss a workaround solution of a machine learning model with combination of cryptographic algorithm which will enhance the effectiveness of the system response to the possible attacks. The other part of the problem, which is hard part to create an intelligence for the organizations for preventing the ransomware attacks with the help of intelligent system password management and intelligent account provisioning. In this paper I elaborate on the machine learning algorithms analysis for the intelligent ransomware detection problem, later part of this paper would be design of the algorithm.**

I. INTRODUCTION

The impact and necessity of information security has increased exponentially over the last few decades as the denial-of-service attacks are increasing, ransomware attacks are increasing, information is being stolen from authenticated data sources, hackers are using more sophisticated and smart methods with help of agile tools for stealing sensitive information. Do small/mid-size/large, government and non-profit

organizations need the security of their system? Yes. They have sensitive user data, employee data, trading data, customer data and other sensitive confidential information stored in office systems. Do common people need the security of their systems at home? Yes. They may have their taxes files, social security card information, bank account details, private pictures, marketing strategy for their small business and many more private things. Some reports clearly shows the motivation behind ransomware is not just money but in favor of some nation's interest, but overall ransomware has affected a broad spectrum of organizations [1].

The large-scale ransomware attacks have increased 135% since 2015 and this has large impact on the performance of individuals [2]. Cryptographic techniques used in the information technology domain to prevent the attacks with the help of complex cryptographic algorithm implementation. Cryptography is used to encrypt the user or organization data and delete the original data from the user system and eventually ask for ransom to recover the hijacked sensitive data.

II. What is the problem being solved?

This scientific model is based on the ardent requirement to protect the system and user data from the novel system attacks in the form of ransom and to threat the user identity data. The identification of suspicious pattern recognition through machine learning algorithm to resolve the major issue of the system attack in the form of ransomware and other type of suspicious activities on user system. The mainstream information security systems are mostly engaged in identity and access management, security policies and standardization of the policies but the undermine problem of resolving the system attacks with the help of cognitive and pattern recognition is remain unanswered.

The main system infections vectors are

1. Spam - unsolicited emails where the malware as an attached file
2. Corrupted web page – files is hacked by malware hacker, the files offered download as substituted for malware
3. Vulnerabilities – hacker delivers the malware to the host of the users operating system and benefits with these vulnerabilities
4. Phishing – fake email and web pages to downloaded to run the malware

With these system vectors the ransomware find the proper path the attack the system to gain the system hosted files in exchange with the ransom.

Figure shows the steps involved in the ransomware detection for the standard ransomware detection mechanism.

Steps involved in the ransomware activity are [1]:

Infection is attached vector in the form of email and reached to the user's system and executed by user. C& C server is contact command and control to obtain or store the encryption key. Then the user's data file is encrypted followed by the extortion money in the form of ransomware.

The main task in the ransomware lifecycle is encryption mechanism of the content of the user files, rendering those files unusable unless the user pays ransom to obtain a decryption key from the hacker. Fast encryption requires a CPU resources, therefore there exists a tradeoff between the longer encryption time and CPU load both will eventually help into ransomware detection [1].

Any form of system hacking in this form can be prevented with the help of implementation of cryptographic algorithm with the machine learning algorithm. Some of the hypothesis of the problems are as describe below.

Hypothesis:

H1: The machine intelligence with machine trained models with the accurate data patterns will help to identity the ransomware attacks

H2: Application security and cryptographic algorithms are good at the application level but for the intelligence to put into the systems need machine learning algorithms

H3: The millions of dollars from the individual to the well-established organizations could be saved with the system intelligence

H4: This research will leverage to develop the applications and could help the monitoring and support vectors in the organization

At present there are many types of software to scan the system and detect the viruses, malwares and ransomware exists but for the prediction of these types of attacks preventions based on the AI and machine learning there is no concrete solution. The approach to analysis and design of the ransomware detection and other

vulnerabilities with the machine learning algorithm is new and has good development scope.

Metrics and cross validation [3]

For the vulnerabilities detection we can use the four common performance indicators

True Positive (TP): Indicates the vulnerabilities or ransomware correctly predicted.

True Negative (TN): Indicates the non-vulnerabilities detected as non-malicious correctly.

False Positive (FP): Indicates the vulnerabilities or ransomware mistakenly predicted.

False Negative (FN): Indicates the vulnerabilities or ransomware incorrectly identified as non-vulnerability.

Accuracy – the number of logs samples that classifier algorithm correctly detects, divided by the number of all ransomware and goldware applications

Accuracy = (TP+TN)/(TP+TN+FP+FN).

Precision – the ratio of predicated ransomware that are accurately labelled as a malware

Precision = (TP)/(TP+FP)

Model for Ransomware detection and prevention

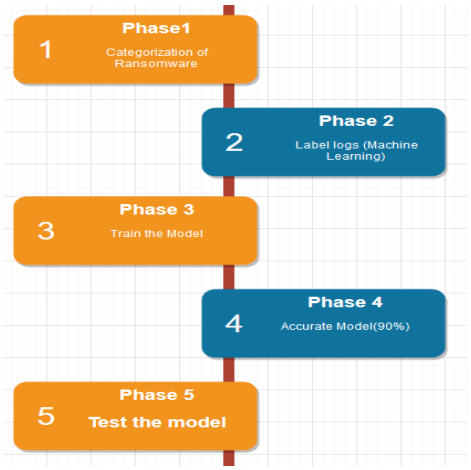


Figure: Model for ransomware Detection and prevention

This proposal will present a solution for the current problem of ransomware detection and prevention using the machine learning algorithm. This application is workable under the circumstances of the large volume of web data log information to analyze the structured and non-structured data from the applications.

III. Why the problem should be solved?

For certain ransomware families the decryption keys can be obtained depending on the key management procedures implemented or the encryption function used, but certain ransomware families the decryption keys are not available in the public domain, in this case the machine learning algorithm techniques can be implemented to decrypt the encrypted data with the help of decryption algorithms solution obtained from the machine learning algorithms this is the hard problem to be resolved.

Machine learning algorithms in the cybersecurity domain has been widely employed and used including the malware detection [3], the approach for the new algorithms is still in demand and many more yet to be published. With the help of the classification techniques on the Neural Network (NN) and supervised leaning the blend of the security algorithms with the machine learning algorithms can work together to resolve the major and common problem ransomware and malware detection.

H1	H2	H3	H4
Machine Learning Models	***Machine learning algorithms***	***Ransom saved through implementation***	***Monitoring time and energy saved***
Dataset required need to be very dynamic and latest	Classifying the vector machine algorithm with 88.2% accuracy	Erebus encrypted 135 Linux servers for 550 Bitcoins [5]	Servers monitoring the attacks
The sandbox env is not sufficient	The model with the predictive analytics	The accuracy in the training data is not evaluated and needs to be evaluated	No specific measures described in the researches from past

Encryption Algorithms used- Both the symmetric and asymmetric key algorithms are used in the different types of malwares and ransomwares [1].

To prevent any form of damage(ransom) to the government and other information sensitive organizations such as healthcare. The password change by machine with the suspicious activities on the systems and notifying the user via email.

IV. Why the problem has not been solved?

Many applications are developed for the information security governance, password reset, single-sing-on but the robust machine learning algorithm for the pattern recognition has not yet developed for these types of the problems. The area of machine learning to solve the more real times problem like intelligent cybercrime detection is still wide open. Many organizations still follow the same traditional monitoring systems and corrective action mechanism for the ransomware detection and prevention. The intelligence is something we need to the push into the systems to recognize the problem ahead of the any form of damage it would cause. One of the papers talks about Text-CNN which achieves the highest accuracy at 0.9890, low false positive rate at 0.03, highest true positive rate at 0.9989 among all other selected classifiers [4].

Table1: Hypothesis and the problems around hypothesis from the literature review.

V. What is the hard part of the problem?

The hard part of the problem is lies in the deterministic objectives of the problem. General objective is to design an algorithm of the ransomware detection and prevention mechanism by analyzing the pattern working on the data pattern and suspicious behavior through tools and interpret the security intelligence.

General Solution approach

Design a model/application which could be used in organizations for the security intelligence

Specific solution approach

Design the unique model/application which can be used to generate the hybrid random cryptographic algorithms to prevent the ransomware attack. The mechanism has two separate parts where the first part is for the preventive attacks and second part is corrective action/s should take by machine to prevent further attacks.

Machine learning model should develop for new detection mechanism and prevention mechanism to provide insider system informatics to the public organization and other who are interested in using this machine learning model.

Hard part of the problem is while we are answering the question from the second hypothesis of my research. The second hypothesis mentioned

Application security and cryptographic algorithms are good at the application level but for the intelligence to put into the systems need machine learning algorithms

Cryptographic recovery from the ransomware with machine learning model:

Files encrypted by ransomware can be recovered if the decryption key can be obtained with the help of trained machine learning and cryptographic algorithm.

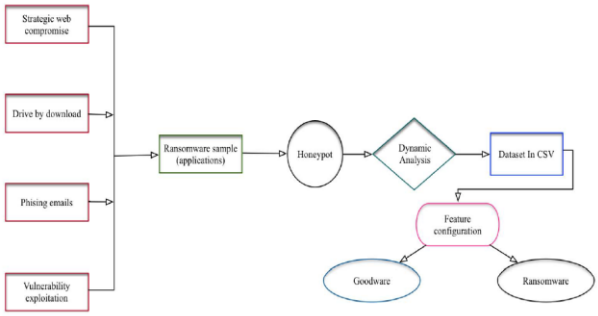
This is the hardest part of the problem as we are designing the model to prediction and identification of the ransomware problem but also for the corrective measure for recovery of the encrypted data by the hacker for ransom.

The cryptographic proposal presented in one of the researches where the intercepted function was used for random numbers generation in the process of session key creation, where the random numbers provided were controlled by ransomware protection software and using the same numbers the session key for the decryption generated, with the help of machine learning algorithm, here the scope for the ransomware decryption key generation by the machine.

VI. What is the solution approach?

Log files provides the relevant and vital information about application and system activities occurring in the specific period. Analyzing log files can provide the valuable insights to the system vulnerabilities and misconfiguration on the system [6]. The logs files analysis with the help of machine learning is the key for most of the vulnerability detection.

The architecture of the system describing the stages in the ransomware data classification problem is as shown below [7]



Algorithm 1: Data cleaning mechanism

Input: Generated data from the logs and other files

Output: clean data

step1: evaluate the sample

step2: remove the samples with bad quality

Algorithm 2:

Input: feed the clean data to the ransomware detection model

Output – the prediction on the data.

Algorithm 3:

Input: the cryptographic algorithms (AES, DES, 3DES)

Output: The blend of the cryptographic algorithms to prevent the further attack.

Literature review suggests that the classification performance on the test data with the highest accuracy achieved with the Text-CNN, and this can be implemented in the new algorithm design.

Below table [4]gives the accuracy and

FRP- false positive rate, TRP – true positive rate

Classifier	Accuracy	FPR	TPR	F_1-score	AUC
Text-CNN	0.9890	0.030	0.9989	0.9796	0.9950
XGB	0.9308	0.023	0.7963	0.8557	0.8869
LDA	0.5048	0.574	0.7698	0.4077	0.6136
Random Forest	0.9348	0.213	0.9861	0.9497	0.8866
Naive Bayes	0.8704	0.250	0.9122	0.7488	0.8457
SVM-linear	0.4420	0.074	0.3587	0.4906	0.8130
SVM-radial	0.7417	0.997	0.9979	0.0061	0.9055

Table 2: Classification performance on the test set. Text-CNN achieves the highest accuracy at 0.989 and low false positive rate at 0.03 among all selected classifiers. XGB performs second best with accuracy at 0.931 and lowest false positive rate at 0.023. All other classifiers either suffer from low accuracy or high false positive rate.

With the text CNN method and cryptographic algorithm in place the model needs to be developed for the ransomware detection.

Blending of machine learning with the cryptographic algorithm for the ransomware detection and further prevention is challenging work.

VII. Conclusions

In this paper, proposed machine learning algorithm can be modeled for the novel ransomware detection and the random number decryption techniques for the new model to break the encryption for saving the ransom. The study also suggested for the classification of the infectious files can be differentiated by the machine based on the model it has trained.

The main problem has structurally divided into sub problems of the identification of the ransomware problems and the design the cryptographic algorithms based on the machine learning to generate the decryption key for the ransom problem.

My ongoing work will be around producing the results and accuracy of the algorithm which I am working on for my research.

VIII. References

[1] D. M. 2. E. M. A. M. I. EDUARDO BERRUETA1, "A Survey on Detection Techniques for Cryptographic Ransomware," *IEEE Access,* vol. Digital Object Identifier 10.1109/ACCESS.2019.2945839, 2019.

[2] M. Lab, "McAfee Labs Threats Report," McAfee, 2016.

[3] A. A. ·. A. D. ·. M. C. ·. K. R. Choo4, "Detecting crypto-ransomware in IoT networks based on energy consumption footprint," open access publication, 2017.

[4] C.-Y. Y. A. P. R. S. Li Chen, "Towards resilient machine learning for ransomware detection," Hillsboro, OR 97124, 2019.

[5] U. Adamu and I. Awan, "Ransomware Prediction Using Supervised Learning Algorithms," *IEEE,* vol. 0.1109/FiCloud.2019.00016, 2019.

[6] V. M. N. P. A. J. T. F. Piyush Nimbalkar, "Semantic Interpretaion of Structured Log Files," *IEEE,* 2016.

[7] U. A. &. I. Awan, "Ransomware Prediction Using Supervised Learning Algorithms," *IEEE,* 2019.

[8] M. H.-A. Juan A. Herrera Silva, "Large Scale Ransomware Detection by Cognitive Security," *IEEE,* pp. 978-1-5386-3894-1, 2017.