INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 4, Issue 1)
Available online at www.ijariit.com

Anomaly Detection Using Machine Learning

Yash Shahani
yash.shahani@ves.ac.in
Vivekanand Education Society's Institute of Management Studies and Research, Mumbai, Maharashtra

Vedant Yadav
vedant.ajaykumar@ves.ac.in
Vivekanand Education Society's Institute of Management Studies and Research, Mumbai, Maharashtra

Aditya Subramanian
aditya.subramanian@ves.ac.in
Vivekanand Education Society's Institute of Management Studies and Research, Mumbai, Maharashtra

Karan Chhabria
karan.chhabria@ves.ac.in
Vivekanand Education Society's Institute of Management Studies and Research, Mumbai, Maharashtra

Vidya Zope
vidya.zope@ves.ac.in
Vivekanand Education Society's Institute of Management Studies and Research, Mumbai, Maharashtra

ABSTRACT

In this day and age of plethora of information, the importance of information security cannot be emphasized enough. Any threat to confidentiality, integrity or availability of information must be taken seriously. Ignoring such threats can have serious consequences, like misappropriation, modification or encryption of data. Vulnerabilities in information security are a tempting target for malwares. Malwares are malicious scripts or software, including computer viruses, worms, Trojan-horses, ransomware, spyware, adware, etc. The traditional way of detecting an advanced malware or threat compromise uses a signature based antivirus. This approach, however, is not foolproof and can be bypassed. The signature based approach relies on a known list of signatures. The list of signatures is not perfect and also does not contain previously unseen malware signatures. The proposed system uses operational intelligence tools and machine learning to monitor usual user behavior. This is done by collecting system activities like event logs, sysinternal, etc. Once the system learns normal behavior patterns, it can detect anomalies that may be caused by malware. Thus, unlike signature based approach, the proposed system can detect previously unseen malwares as well.

Keywords: *Malware, Machine Learning, Antivirus, Operational Intelligence Tool.*

1. INTRODUCTION

The digital world today is ridden with a myriad of different harmful elements. The development of sophisticated, protocol based environment has regulated the activities on the internet. That being said, this kind of sophistication has also brought about an increase in the number of attacks such as those based on anomalies like malware, adware, etc.. These attacks have nullified the idea that basic netiquette is enough for a safe traversal through the digital world. Though netiquette ensures that unnecessary conflicts are avoided, still there are many unscrupulous participants in the digital world who cannot be controlled. This set of individuals or rather group of individuals create and spread anomalies like malware, ransomware, adware, etc. over the internet and/or remote systems. Traditional methodologies like antivirus, IDS, etc. do exist for defence against such attacks but the attackers are also getting smarter day by day. The proposed project aims to provide an insight into a newer approach to increase the effectiveness of techniques used to curb these attacks. Machine learning algorithms and Big data analytics will be used for implementing the said approach.

2. RELATED WORK

This section provides the research work carried out for understanding the existing practices and the systems in place for malware detection using machine learning. It also involves listing of the Machine Learning algorithms that have been used in existing implementations.

Fraley et al. [3] suggest the use of machine learning for detection and mitigation of security events and alerts over the traditional malware detection systems, which are not able to handle new and unknown threats and rely more on security personnel for analysis and mitigation. The key points of this solution is to classify alerts using Deep Neural Networks (DNN), utilizing Machine Learning to provide rapid analysis and recognize threats and highlight & generate hidden threat patterns. The proposed solution first identifies the threat, then classifies it based on attack behavior & effects on the system and finally provides solutions based on type of alerts. The solution focuses on Machine Learning to accurately predict the actions of a security analyst, given a dataset.

Dragos et al. [2] introduce a framework to distinguish between malware and clean files by using a simple multi-stage combination (cascade) of different versions of the perceptron algorithm. They use three data sets: a training dataset, a test dataset, and a "scale-up" dataset to minimize the number of false positives.

Markel et al. [6] proposed an approach that primarily learns from metadata, mostly contained in the headers of executable files. The experiments indicate that executable file metadata is highly discriminative between malware and benign software. They also employ various machine learning methods, finding that Decision Tree classifiers outperform Logistic Regression and Naive Bayes in this setting.

Gandotra et al. [1] have proposed a 6 step procedure for comparison of different ML algorithms for measuring their degree of efficiency: Data Acquisition, Automated malware analysis, Feature Extraction, Feature Selection, Classification, Evaluation and Validation. They compared the working of various algorithms when given 18 features (before feature selection) and 6 selected features (after feature selection).

3. PROPOSED METHODOLOGY

This section illustrates the proposed system for detection of anomalies and User Behaviour Analysis.

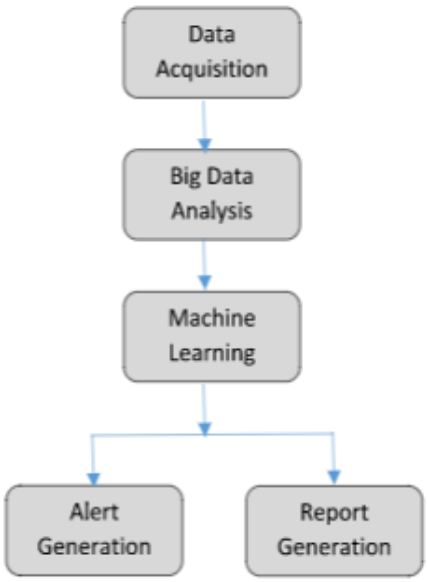

Fig.1 Flow of Proposed Methodology

A. Data Acquisition

The data is collected from system logs as well as user logs. It is collected through various sources and from local as well as remote hosts. Splunk operational intelligence tool provides software for indexing of logs or any form of machine data. This includes structured, unstructured as well as complex multi-line logs. The tool allows collecting, storing, indexing, searching, correlating, visualization, analysis and report generation on any machine data to identify and solve operational and security issues.

B. Big Data Analytics

After the data is collected from various sources big data analysis is used. This acts as user behaviour fingerprint. We use splunk application aster and hunk which enables to detect pattern behaviour corresponding to particular user. This is also used to generate sigma reports.

C. Machine Learning

Splunk enterprise has Machine learning toolkit application which has built-in algorithms. R and python scripts can also be used for user defined algorithm. Classified as predictive numeric and categorical fields, detect numeric and categorical outlier, forecast time series and cluster events. This is used to detect anomaly on basis of pattern behaviour of user. Clustering algorithm like K-mean and Spectral Clustering is suitable for user behaviour fingerprinting and for threat detection, numeric outliers and categorical field prediction algorithms like logistic regression, Gaussian NB, Decision tree classifier are suitable.

D. Alert Generation

Splunk provides with a feature of alerts. One can generate alerts for a particular event or when a threshold value is matched. This alert can also be sent via email.

E. Report Generation

Splunk provides a built-in report generation utilities. Different graphs and charts can be plotted. It also provides wide range of flexibility.

4. SYSTEM DESIGN

This section provides the block design of the proposed system with consideration of a suitable use case. The different modules have been explained in detail.

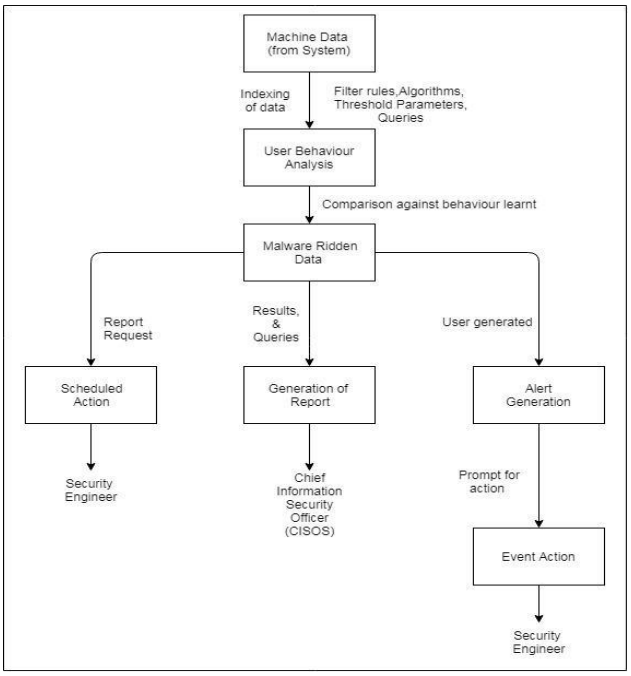

Fig.2 System Design

A. Machine Data

This involves choosing a proper source for collection of logs. This can include network ports, firewalls, file systems, etc.. The data that is collected in this module is indexed. Indexing is done with respect to time stamps in general. This data can be queried, filtered, etc. using commands built into Splunk.

B. User Behaviour Analysis

A combination of Big Data Analytics and Machine Learning is done to learn the normal user behavior in safe state i.e in the absence of anomalies such as malware, etc..

C. Malware Ridden Data

The data set containing anomalies such as malware is processed and compared with the normal user behavior learnt in the previous module. Such comparison will help in detecting any anomalies against normal behavior.

D. Scheduled Action

When considering a use case with actors such as security engineer, system analyst and Chief Security Officer, a periodic analysis of the system will be done in order to monitor current status of the system.

E. Generation of Report

A report is generated which includes information about the number of anomalies detected and time taken for complete processing.

F. Alert Generation

In the event of detection of malware in the current data set, a user-generated alert can be created. This will be useful for commencing mitigation strategies in a timely manner.

G. Event Action

This involves decision on the actions to be taken in the event of reception of user-generated alert.

5. CONCLUSION

The review of current systems indicates that Machine Learning approach is being preferred over traditional approach for malware detection. Comparison of Traditional software and ML based systems for detection of known malware is not yet explored extensively. Major drawback in ML based detection techniques are - 1. Build Time, 2. Learning Time. A common issue with any malware detection software is the use of polymorphism in the script written by malware coders.

Based on the comparison results of various algorithms as seen in related works, it is evident that Linear Regression, Logistics Model Tree, K-Mean and Random Forest are the algorithms giving maximum precision for Malware classification / detection. This list may be appended or modified based on factors such as compatibility of individual algorithms with the operational intelligence tool.

6. REFERENCES

[1] Ekta Gandotra, Divya Bansal, Sanjeev Sofat, "Zero-Day Malware Detection", IEEE 2017 *Patna, India IEEE Sixth Int. Symposium on Embedded Computing and System Design., 2016*

[2] Dragos¸ Gavrilut , Mihai Cimpoesu, Dan Anton, Liviu Ciortuz, "Malware Detection Using Machine Learning" IEEE 2009

[3] James B. Fraley, Dr. James Cannady, "The Promise of Machine Learning in Cybersecurity". IEEE 2017

[4] James B. Fraley, Dr. James Cannady, "Enhanced Detection Of Malicious Software". IEEE 2016

[5] Arshi Dhammi, Maninder Singh, "Behaviour Analysis of malware using Machine Learning", IEEE 2015

[6] Zane Markel,Michael Bilzor, "Building a machine learning classifier for malware detection", IEEE 2015

[7] Weiwei Hu, Ying Tan, "On the robustness of machine learning based malware detection algorithms", IEEE 2017

[8] Splunk white paper: SPLUNK SECURITY USE CASE DETECTING UNKNOWN MALWARE AND RANSOMWARE.

