Real-Time Framework for Malware Detection Using Machine Learning Technique

Sharma Divya Mukesh[1(✉)], Jigar A. Raval[2], and Hardik Upadhyay[3]

[1] GTU-PG School, Ahmedabad, Gujarat, India
sharmadivya666@gmail.com
[2] Physical Research Laboratory, Ahmedabad, Gujarat, India
jigar@prl.res.in
[3] GPERI, Mehsana, Gujarat, India
hardik31385@gmail.com

Abstract. In this epoch, current web world where peoples groups are associated through correspondence channel and the majority of their information is facilitated on the web associated assets. Thusly the security is the significant concern of this internet community to protect the resources and to ensure the assets and the information facilitated on these networks. In current trends, the greater part of the end client are depending on the end security items, for example, Intrusion detection system, firewall, Anti-viruses etc. In this paper, we propose a machine learning based architecture to distinguish existing and recently developing malware by utilizing network and transport layer traffic features. This paper influences the precision of Semi-supervised learning in identifying new malware classes. We show the adequacy of the framework utilizing genuine network traces. Amid this research, we will execute and design the proactive network security mechanism which will gather the malware traces. Assist those gathered malware traces can be utilized to fortify the signature based discovery mechanism.

Keywords: Malware detection · Semi-supervised algorithm · ClamAV · Machine learning

1 Introduction

1.1 Malware Discovery from the Linux System

Malwares are a typical channel by which crackers encourage cybercrime. It has turned into a weapon whose utilization meets different dangers defied by security analysts over the globe. The crackers continue improving the intricacy of malcode to fulfill their notorious aims. Malware is a nonexclusive term to mean a wide range of undesirable programming (e.g., backdoors, virus, worms, spyware and Trojans) [5]. Various assaults made by the malware have represented a noteworthy security risk to PC clients. In this manner, malware identification is one of the computer security points that are of incredible intrigue.

As of now, the most huge line of guard against malware is hostile to infection programming items, for example, Kingsoft's Antivirus, Dr. Web and Norton.

S.C. Satapathy and A. Joshi (eds.), *Information and Communication Technology for Intelligent Systems (ICTIS 2017) - Volume 1*, Smart Innovation, Systems and Technologies 83, DOI 10.1007/978-3-319-63673-3_21

These generally utilized malware detection software apparatuses essentially utilize signature-based strategies to perceive dangers. Signature is basically a short series of bytes that are special for each existing malware so that future instances of it can be effectively ordered with a less error rate. In any case, this exemplary signature-based based strategy dependably neglects to identify variations of known malware or already obscure malware, in light of the fact that the malware essayists dependably embrace methods like obfuscation to bypass these signatures.

Various gadgets are running Linux because of its flexibility and open source nature. This has made Linux stage the objective for malware assaults, so it gets to be distinctly critical to identify and dissect the Linux malware. Today, there is have to break down Linux malwares in a mechanized approach to comprehend its capacities [3].

1.2 Commitment of Research Paper

With a specific end goal to manufacture compelling, automatic, and interpretable classifiers for malware recognition from the substantial, confused and unlabeled list, we need to address the accompanying difficulties:

- **How to develop a successful classifier to distinguish malware from the unlabeled dataset?** How to make the classifier less touchy to the lopsidedness and perform well for the vast and muddled dataset?
- **How to make the classifier interpretable?** The classifier ought to produce learning or patterns that are simple for the malware investigators to comprehend and interpret.
- **How to productively assemble the classifiers?** In our application, we have an aggregate of 5000 labeled file samples from ClamAV that are accessible for training: half of them are malware tests and the other half are amiable record tests. Testing strategies are expected to construct classifiers for such a vast information gathering to keep away from over-fitting and accomplish extraordinary viability and additionally high proficiency. In any case, the decisions of the class circulation and the extent of the training data in the inspecting technique for malware detection are not inconsequential and require careful examination.

In this paper, we depict our examination push to address the above difficulties. To address the interpretability issue, we fabricate various leveled classifiers since they can produce decides that are simple for malware investigators to comprehend and translate.

1.3 Content of the Paper

The straggling leftovers of this paper is dealt with as takes after. Sections 2 discuss about the machine learning information, Sect. 3 exhibits the outline of malware recognition framework architecture and Sect. 4 examines the related work. In Sect. 5, we portray the simulation study of malware detection framework. At long last, Sect. 6 concludes.

2 Machine Learning Information

Machine learning is a method for analysis of statistics that automates illustrative model constructing. Making use of calculations that iteratively benefit from facts, machine mastering allows systems to find concealed bits of statistics without being unequivocally tweaked wherein to look [10].

There are different usages of machine learning. It's very to recognize how much machine learning has finished in genuine applications. Machine learning is consistently associated in the disconnected training phase. Accordingly machine learning is used to improve the applications, for example, face recognition, face detection, speech recognition, genetics, image classification, weather forecast etc. Machine learning is associated in malware detection & classification to enhance the accuracy of the malware detection rate estimate. Machine learning makes it decently less requesting to make complex programming systems without much effort on the human side.

Underneath specified are essential six assignments required in Machine Learning process.

- **Classification** – It will categories new information in predefined trainingvia locating prescient getting to know characteristic.
- **Clustering** – Identify likeness in information and shape gatherings or groups
- **Summarization** – Locating a minimized illustration for a hard and fast or subset of records
- **Regression** – Discovering prescient learning capacity, which models information with the slightest mistake.
- **Dependency modeling** – Locating dating between elements or their features in a dataset.
- **Anomaly recognition** – Identify the most critical changes or mistakes in information set.

Data mining process utilizes framework acing calculations relying upon whether the class names are accommodated becoming more acquainted with, these calculations might be isolated into classifications supervised learning or unsupervised becoming acquainted with.

3 System Architecture

The inspiration to propose this approach is, about sixty percentage interruption and security infringement are inside the association. The security reports created analyser indicate obviously. If the traffic is separated into various classification astute, it will be more complex for further examination too. This building square would be same for open system traffic and authoritative. It is half breed behaviour and intrusion detection framework.

There are basic three phases of proposed architecture. This three phases are imagined as key segment of an expansive end security framework that screens the system stream and choosing whether it is malignant or amiable. Figure 1, demonstrate the proposed

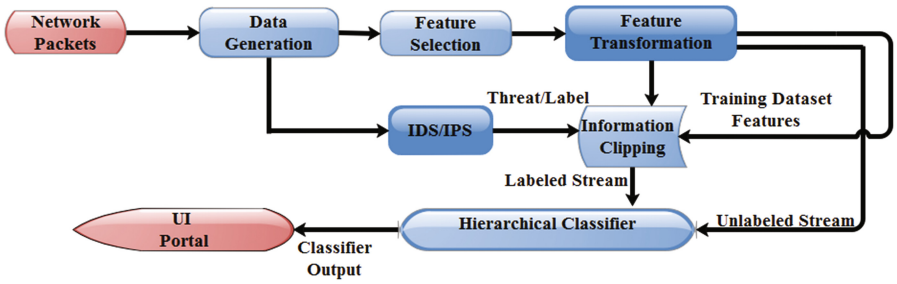

Fig. 1. System architecture

architecture, which consist three noteworthy segments: **Signature Based Malware Detection, Behaviour Based Malware Detection and Machine Learning Phase**

3.1 Phases of Proposed Architecture

First Phase: Signature Based Malware Detection
This module would accumulates every one of the packets from the network and match the approaching system activity with intrusion detection system signature for instance, approaching movement is tried on Clam-AV open source antivirus having different malware signatures (Fig. 2).

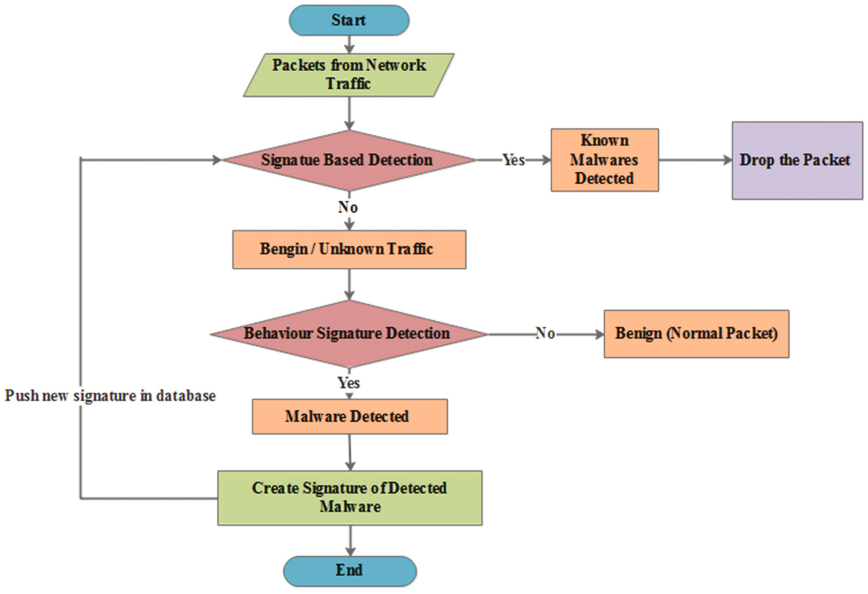

Fig. 2. System architecture flowchart

Second Phase: Behaviour Based Malware Detection
Intrusion detection and intrusion prevention system would perform DPI of packets after that it would names the stream conclude that it belongs to some existing threats On the off chance that the approaching traces will sidestep the Clam-AV then will consider that network traces for behavior based detection in which include feature selection is accomplished for incoming network traffic. That traffic would be consider as level to classifier in hierarchical class fifer phase of system architecture
Packet network packets and collect the value of network layer 3 and layer 4 features and gathering them in light of same stream. Features like packet per flow, bytes per second, packet inter arrival times and payload size.

Third Phase: Machine Learning Phase

Step 1: Input the information which is network packets.
Step 2: Parsing of network packet to collect feature selection.
Step 3: Construct hyper sphere based on view of classes saw in preparing set.
Step 4: Compute distance of test instance focal point of each hyper sphere.
Step 5: If the test instance lie outside the hyper sphere then class is anticipated as unknown malware sample.
Step 6: If the test instance is variant of existing malware then, it will live just inside a solitary hyper sphere.
Step 7: Compare the distance profile of the instances to all hyper sphere.

This structure utilizes Classifier work with Class-1 Support vector machine algorithm. This is semi supervised machine leaning algorithm, where each model invests huge energy in requesting cases from a specific malware class. This would have an ability to perceive new malware besides.

4 Related Work

An essential evaluation of the work has been completed as such far for malware recognition to show how the present review identified with what has as of now been finished. In light of the investigation of the security gadgets and their working capacity, it is important to put the proactive security instrument rather than the signature based techniques such as firewall, Network Intrusion Detection System (NIDS), etc. Since the identification algorithm of these network intrusion detection system depends on the how signature based antivirus apparatuses distinguish malevolent exercises. All these signature based gadgets are depend with respect to the pre-decided lead sets as assault database that as of now been distinguished and recorded. This leaves in the condition of known as obscure assaults. Thusly they won't have the capacity to distinguish the obscure assaults which bargain the PC framework. In view of the previously mentioned methods, there is a necessity to set up a proactive based security tried that would be capable review the whole system and in the meantime gather the assault follows which is known and obscure to them. The unknown attack traces can be utilized to fortify the network intrusion detection system, prevent intrusions from exposing and exploiting the vulnerabilities of the said network. One of such proactive system security

instrument is as honeypots. Since honeypots are misleading frameworks, it will be exceptionally helpful secluded from everything the genuine estimation of the information that disregard the system.

In this section, we cover research efforts claim to detect malware variants. We group the research technique based into two groups'. Namely, they are signature-based, behavior-based.

4.1 Behavior Based Detection

API calls and in addition System calls are likewise utilized by for malware detection utilizing behaviour based malware detection approach (Mamoun Alazab et al. 2012, Hisham Shehata Galal, Yousef Bassyouni Mahdy et al. 2015, Stavros D. Nikolopoulos, Iosif Polenakis et al. 2016), perform well for classification by concept of dynamic analysis technique to extract API details of latest malware dataset inside controlled environment, API hooking technique is used means to trace the API calls invoked my OS. That traces will function as feature that is called as action. All this actions are then classified based on various algorithms. This model fails in front of malware samples that checks for the existence of virtual machine artifacts. For future improvement is needed the behaviour of malware actions, data flow dependence can be used which will give more insight for malware identification. While in the case of malware detection by considering system calls disadvantage is that grouping of system call to system calls groups leads to loss of information leading to higher rate false positive and Unknown malware to known malware failure rate is 83.42% i.e. classification rate [2, 8, 10].

4.2 Signature Based Detection

Keeping in mind the end goal to conquer the impediment of the generally applied used signature primarily based malware detection method, facts mining and system getting to know approaches are proposed for malware detection (Sachin Jain, Yogesh Kumar Meena et al. 2011; Han-Wei Hsiao, Deng-Neng Chen, TsungJu Wu et al. 2011), it is shown vindictive internet site detection, and it utilizes the spatial- temporal aggregating factors way to deal with detection system from Net Flow data For arrangement of packets it utilizes the induction learning technique on the way to differentiate between malicious and normal traffic pattern. Hence through the utilization of various variables it had been presumed that Net flow variable will produce more awful detection accuracy rate, combination on spatial and temporal variable may have greater accuracy rate [1, 7, 9]

In spite of the fact that Semi-supervised classification strategy performs well on extensive information gathering, it should be further explored for managing class imbalanced issue.

5 Simulation Study

The experiment of proposed system architecture has been portrayed in this section. Underneath Fig. 3, demonstrates the network architecture setup to implement propose malware detection architecture.

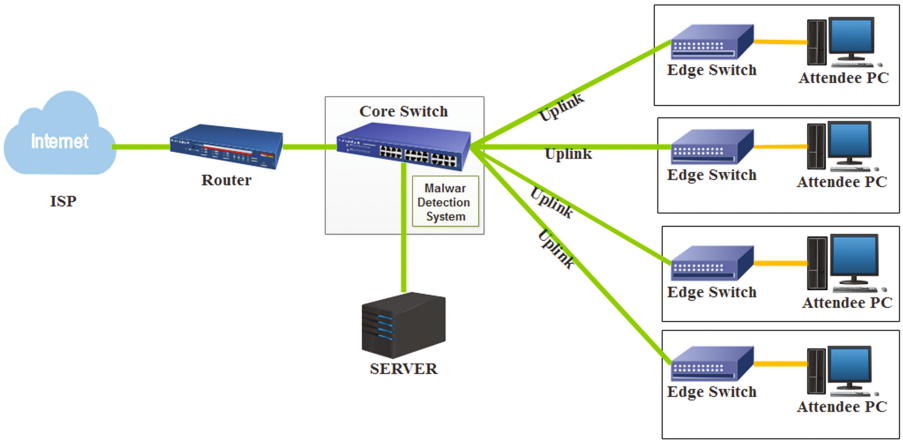

Fig. 3. Network architecture

We have utilized ClamAV for implementing signature based malware detection module which have essentially three database in which signature are stored in hex patterns, MD5 patterns and LDB signatures this are overhaul by utilizing freshClam command after establishment of ClamAV in the system.
The following are the steps which demonstrates to setup ClamAV:

Step 1: sudo apt-get install clamav
Step 2: sudo freshclam
Step 3: sudo nano freshclam.conf

Figure 4, demonstrates how existing signature are been utilized to coordinate the malware yet by utilizing this strategy zero-day malware whose signature doesn't not exist, can't been identified so its testing errand to beat this issue, by utilizing behavior based technique it would be supportively to detect new malware and variants of new malware.

```
root@kali:~/ForensicsTools/clamav/sigs/test# clamscan -d sig.ndb ../../malwarez/sploit.exe
../../malwarez/sploit.exe: Exploit.RingZ.UNOFFICIAL FOUND

----------- SCAN SUMMARY -----------
Known viruses: 1
Engine version: 0.97.8
Scanned directories: 0
Scanned files: 1
Infected files: 1
Data scanned: 0.05 MB
Data read: 0.05 MB (ratio 1.00:1)
Time: 0.014 sec (0 m 0 s)
```

Fig. 4. Scan Output from ClamAV

So as to overcome the challenge of signature based malware detection will use behavior based malware detection approach that is by considering feature selection and transformation of network traces which is unlabeled data set. This feature are consider as testing data. When the malicious traffic is encountered then it is being transferred to real database server then it will drop the request if signature matches with the existing signatures of malware. If the match is not found then the traffic is redirected towards the honeypot server. In the honeypot, the request will be executed and result would be screen, if any pernicious traffic is experienced here, then new signature is generated and that new signature will be interface with the Clam-AV or any IDS system. underneath figures shows implementation of behavior based malware. And It has been identified utilizingGlastopfhoneypot when remote file inclusion attack is performed on live website. In this we had upload lab.php file which is fundamentally backdoor Trojan by which we can bypass authorization. This malware will work only if the web framework have file inclusion vulnerability. Once malware identified then will parse specific network packet to identify behavior changes like client to server load or vice versa. In Fig. 5 shows the implementation of behavior based malware detection approach

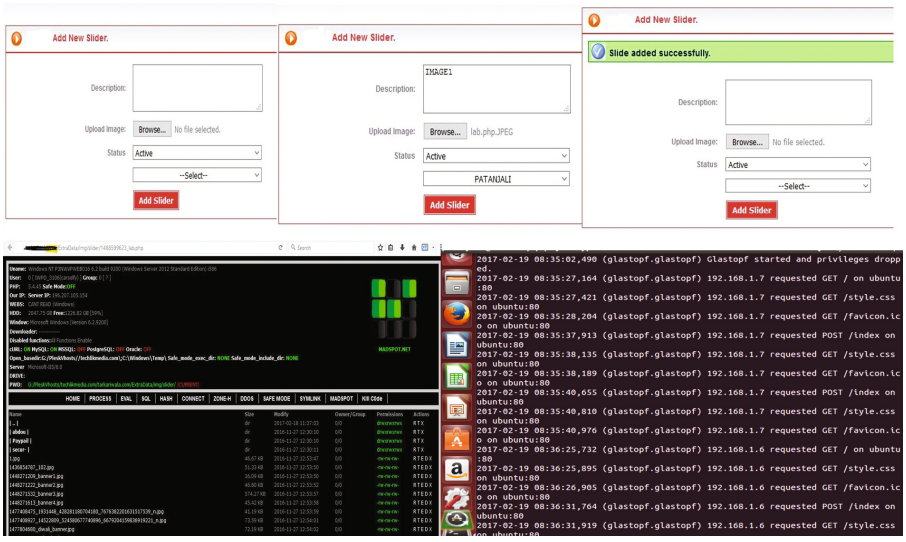

Fig. 5. Remote file inclusion attack using malware

In Figs. 6, 7, graph shows the comparisons of existing and proposed system.

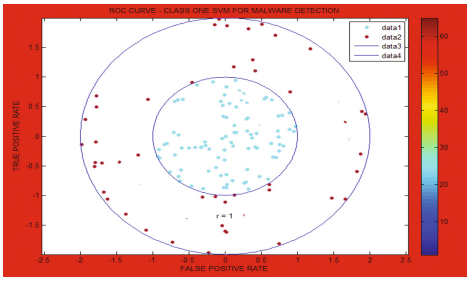

Fig. 6. ROC curve analysis

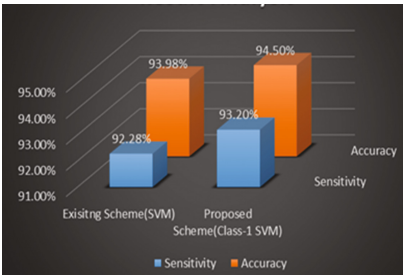

Fig. 7. Result analysis

6 Conclusion and Future Scope

To identify malicious packets, this paper present a new machine learning approach which depend on classification methodologies for detecting malware behavior. Features discover from the network and transport layer network stream attributes. In spite of the fact that the components are stronger for encrypted payload, it would have different concern which would hamper the appropriateness of extended complex learning algorithms. Besides, the supervised algorithm tough for distinguishing new variants of malware. Our proposed method addresses those difficulties and recognize streams of present and novel malwares with excessive accuracy. At last, we display a novel adjustment of class one Support vector machine to distinguish unheard-of malware.

Meanwhile, there are still many works to do in future. Currently, focuses on capturing only non-encrypted traffic. For destiny work, we strive to extend the method to SSL network traces and additionally do on line getting to know placing. Likewise, whilst the amount of training turns out to be significantly extensive, the forecast step is expensive due to the huge wide variety of hyper spheres that ought to be tried. To deal with this trouble, we plan to build up a revolutionary multiclass learning method.

References

1. Chuan, L.L., et al.: Design and development of a new scanning core engine for malware detection. In: 2012 18th Asia-Pacific Conference on Communications (APCC). IEEE (2012)
2. Hsiao, H.W., Chen, D.N., Wu, T.J.: Detecting hiding malicious website using network traffic mining approach. In: 2010 2nd International Conference on Education Technology and Computer, vol. 5. IEEE (2010)
3. Sochor, T., Zuzcak, M.: High-interaction linux honeypot architecture in recent perspective. In: International Conference on Computer Networks. Springer International Publishing (2016)
4. Bazrafshan, Z., et al.: A survey on heuristic malware detection techniques. In: 2013 5th Conference on Information and Knowledge Technology (IKT). IEEE (2013)
5. Galal, H.S., Mahdy, Y.B., Atiea, M.A.: Behavior-based features model for malware detection. J. Comput. Virol. Hacking Tech. **12**(2), 59–67 (2016)

6. Saeed, I.A., Selamat, A., Abuagoub, A.M.: A survey on malware and malware detection systems. Int. J. Comput. Appl. **67**(16), 25–31 (2013)
7. Ahmed, Irfan, Lhee, Kyung-suk: Classification of packet contents for malware detection. J. Comput. Virol. **7**(4), 279–295 (2011)
8. Anderson, B., et al.: Graph-based malware detection using dynamic analysis. J. Comput. Virol. **7**(4), 247–258 (2011)
9. Jain, S., Meena, Y.K.: Byte level n–gram analysis for malware detection. Comput. Netw. Intell. Comput. **157**, 51–59 (2011)
10. Alazab, M., et al.: Zero-day malware detection based on supervised learning algorithms of API call signatures. In: *Proceedings of the Ninth Australasian Data Mining Conference-Volume 121*. Australian Computer Society, Inc. (2011)
11. Nikolopoulos, S.D., Polenakis, I.: A graph-based model for malware detection and classification using system-call groups. J. Comput. Virol. Hacking Tech **13**(1), 1–18 (2016)