Malware Detection on Server using Distributed Machine Learning

Usman Aijaz N, Anisha Patra, Ayesha Siddiq S, Bichitra Chatterjee, Mehfooz Ghiyas Khan

Department of Computer Science Engineering, Brindavan College of Engineering, Bengaluru, India

***Abstract:* Malware has continued to develop at a disturbing rate despite on-going reduction efforts. This has been considerably more pervasive on web servers, where server computing is an increasingly popular platform for both industry and buyers. As of late, another age of malware families has developed with advanced evasion abilities which make them substantially difficult to identify utilizing ordinary techniques. On one hand, the popularity of worldwide use of Internet absorbs attention of most engineers for delivering their applications. The expanded number of applications, on the other hand, prepares an appropriate prone for a few users to create various types of malware and include them in the market or in outsider markets as sheltered applications. This paper proposes and explores a machine learning based characterization approach for identification of malware and utilizes distributed Support Vector Machine (SVM) algorithm keeping in mind the end goal to detect malicious software(malware) in server computing platform using malicious and benign records.**

***Keywords:* Malware detection; Server computing; Machine learning; SVM**

I. INTRODUCTION

Malware [1,3], short for malicious software, disturbs computer activities, assembles sensitive data, as well as gets control to private computer systems. Malwares take client data, make premium calls, and send SMS ad spam's without the client's authorization. And as the number of internet users increases the malware threats also increase. Therefore it becomes important to detect the malware before it might lead to any vulnerability

We realize that the majority of downloads are done from the server for our everyday activities. The functionality of a computer server is to store, recover and send computer documents and information to different computers on a network. On a bigger scale, the worldwide computer network known as the Internet relies upon countless servers situated far and wide. The server services are relied upon to be critical in nature, as they are prominent within the private, public and commercial domains. Thus, the security and resilience are progressively vital aspects.

And as mentioned earlier, malwares can be harmful to the clients or the users connected to the server. It is evident that there is a need for detecting these malware, and to stop them from kidnapping our privacy and disturbing our lives.

Here in this paper we utilize an online server anomaly detection approach [2], including committed recognition segments. Anomaly detection also known as outlier decision is the identification of items, events which do not conform to an expected pattern or other items in a dataset. Online anomaly detection has the benefit that it can permit experts to carry out corrective actions as soon as the anomaly is occurred in the sequence data.

For this purpose we are utilizing the machine learning approach, which is a field of computer science that enables the computers to "learn" information, without being expressly programmed. The goal here is to detect anomalies, that is, actions that deviate from the normal behaviour of the legitimate user. As already pointed out, such actions may arise for a number of reasons, including malware, illegal use of the device and so forth.

II. RELATED WORK

Malware detection was approached with many other techniques. The methods used are signature based, behavioral based and heuristic based. Signature based attempts to model the malicious behavior of malware and uses this model in the detection of malware. The system detects intrusions by observing events and identifying patterns which match the signatures of unknown attacks. Behavioral based evaluates the malware based on its actions. It observes and evaluates in context every line of code executed by the malware. They analyze all requests to access specific files, processes, connections, or services. If malware determines it’s running in a sandbox, it’ll attempt to avoid detection by curtailing malicious activities. It also takes time to analyze the behavior of an object. Heuristic based is capable of detecting previously unknown viruses as well as variants of new viruses. It performs the function by executing the programming commands of a script (questionable program) within a specialized virtual machine.

These methods of malware detection [1] had a lot of disadvantages such as, with a lot of datasets to train and test it requires extensive evaluation which cannot be done

© ΠCES 2018. All right reserved.

KITE was held at Brindavan College of Engineering, Bengaluru, India on 4th May, 2018.

with the above methods because they consume a lot of time for n number of datas. Hosts that are subjected to large amounts of datasets the detection system can have a difficult time inspecting every single data that comes in contact. System can suffer a substantial performance slow down if not properly equipped with necessary hardware to keep up with the demands. And the effectiveness is fairly low regarding the accuracy. But to a far more extent, if we see the upside of these methods they were used efficiently to come up with the solutions of detection until a new approach was found, which has the capacity for improved detection accuracy, fast processing and real-time predictions and much more, this approach is called Machine Learning.

And in our paper, we are using machine learning associated with the server computing [7,18] for the detection and identification of malware. It has the advantage of easier detection of malicious documents or data which are uploaded on the server, with the server presenting a number of unique security issues. And the client or user gets prior information if he/she attempts to download the malware file.

The framework portrays a multi-layered solution of the server nodes and considers how malware identification can be conveyed to every node. We're utilizing a distributed machine learning detection approach that utilizes the Support Vector Machine (SVM) [15] algorithm. SVM is defined as a supervised learning model associated with learning algorithms that analyze the data used for classification and regression analysis. It also demonstrates the effectiveness of detection under different malware types, as there are many different malware families that go undetected as they conceal the malicious payload they can't be promptly spotted in the wild.

III. MACHINE LEARNING AND ITS TYPES

Machine learning [2,16] is a type of Artificial Intelligence where the machine learns from its code, we write the program once and when the machine encounters another problem, it should not be programmed again. It changes the code according to the new scenario's it discovers. Machine learning can be categorized into major groups as supervised, unsupervised machine learning and reinforcement learning[10,12] as shown in Fig. 1. These groups represent how the learning method works.

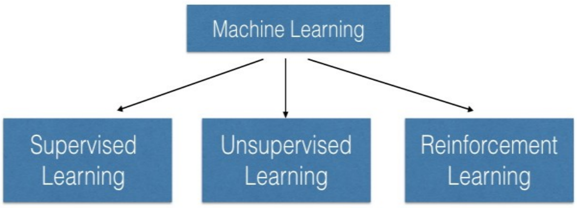

Fig 1. Machine Learning Types

Supervised learning: It is a machine learning algorithm[9,11] that uses a known dataset to make predictions. The dataset includes input data and response values. From it, this algorithm seeks to build a model that can make predictions of the response values for new dataset.

Unsupervised learning: It is a machine learning algorithm used to draw interfaces from datasets consisting of input data without labelled responses. It finds a pattern or structure behind those inputs.

Reinforcement learning: It is an area of machine learning concerned with how software agents take action in environments so as to maximize the reward.

Classification: Classification algorithm is a part of supervised learning, used to classify records. It is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y) [10].The output variables are often called labels or categories. The mapping function predicts the class or category for a given observation. For example, a file or a document can be classified as belonging to one of two classes: "spam" and "not spam" or "malicious" and "benign".

Regression: It is the task of approximating a mapping function (f) from input variables (X) to a continuous output variable (y). A continuous output variable is a real-value, such as an integer or floating point value. These are often quantities, such as amounts and sizes.

Support Vector Machine (SVM): It is a supervised learning model associated with learning algorithms that analyze the data used for classification and regression analysis. It is the most used algorithm for classification tasks and suited for extreme cases. It best segregates two classes. SVM [12] draws a decision boundary between two classes called as the hyperplane. And with this hyperplane comes up with the decision of prediction of a particular parameter as it falls in which class.

IV. METHODOLOGY

We chose to develop detecting malware [1] on server using distributed machine learning rather than the conventional methods like the signature, heuristic and behavioural as it satisfies the user requirements of accurate predictions with large amount of datasets, timely prediction and fast processing. Associating malware detection with server computing enables us to get unique security issues. We demonstrate that our scheme can reach a high detection accuracy of over 90% whilst detecting various types of malware and DoS attacks. Furthermore, we evaluate the merits of considering not only system-level data, but also network-level data depending on the attack type. The processing takes place as follows: where we are considering a service provider, server and end-user entities.

A. Service Provider

It is an entity that provides services to the end-users [18] authorizing it from the server. It consists of the users, document ranking, with the document name, domain and sub-domain. Here we add content, view its details, list all search history and user's, it also auto recommends the

© IICES 2018. All right reserved.

document to users and measures the expectation loss if content is not matched. The service provider authorizes the admin and registers the user. The procedure starts with the admin registering itself and then uploading the documents. It does so with the RSA key, by encrypting the document and uploading to the server. If the admin wishes to upload the documents at a later time it simply has to login with his/her credentials to access the service provider.

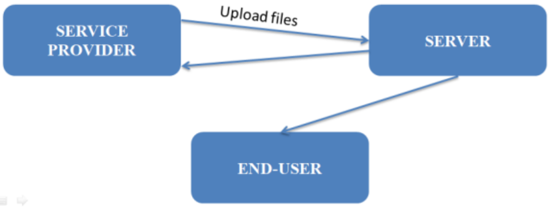

Fig 2. Server Computing

B. Server

It provides the functionality to its clients [18] when there is a request. The tasks of the server are, it views all the files uploaded by the service provider, authorizes the service provider and the end-user, processes the query coming from the user for document request etc. The operation of the server is as follows, it registers the admin and the end-user, with their credentials.

Associating with the service provider, the operation of the server is, it confirms the registration of admin and the request of uploading the documents. It diagnoses the malicious users who upload malware files, finds the percentage of the malware. Associating with the end-user, the operation of the server is, it confirms the registration request and query search. When a user requests for a document with the title the server looks for the presence of the document and analyzes it for the malware by using the machine learning SVM approach. It the requested document happens to be a malware file, the server sends a alert message to the user confirming it as a malicious document. If the file is clean/benign, it lets the user to download it.

C. End-User

It is the one who is intended to ultimately use the product. The tasks of end-user are only to search for documents and view its search history. Its operation is as follows: it registers itself with the server and writes query to request for a particular document. And also registers the secret key to decrypt the document which was encrypted by the service provider. The end-user gets the alert messages from the server if the requested document in a malicious file. This is the task of an end-user.

D. SVM

The machine learning analysis [16,17] of the SVM is done at the server stage of implementation, where the requested document is analyzed for any malware. SVM is the best suited algorithm for this purpose as it is memory efficient and can be used for higher-dimensional spaces. It is used in many applications such as medical imaging, image interpolation, fault diagnosis etc.

The operation of SVM [12] is as follows: the request for a particular file is received and for machine learning to predict whether it is a malicious or benign file it places the files on a linear graphical plane. These files are arranged based on the parameters which classify them. For instance, the parameters also known as feature can be such as free_area_cache, truncate_count, last_interval, shared_vm, exec_vm and so on. In the Fig.3 we see the classification of the two different classes based on the parameters, they are the benign and malicious.

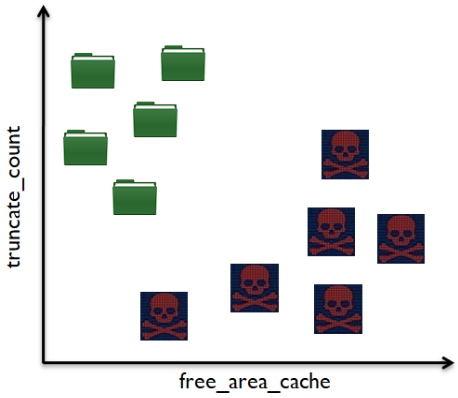

Fig 3. Classification of files based on parameters (features)

After the classification of the files we draw a decision boundary between them for optimized classification called as the hyperplane [13]. The hyperplane segregates the two classes in the best way possible, and the margins are drawn beside it where in the extreme datapoints fall. They are called as the support vectors. These support vectors are done extensive evaluation for predicting that it's a malware or benign file. Fig 4. shows the hyperplane with support vectors.

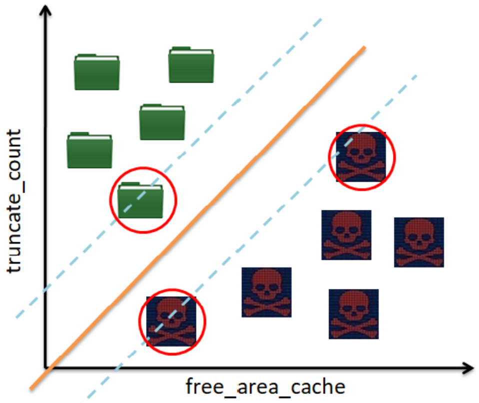

Fig 4. Hyperplane and support vectors

The two thumb rules of SVM [9] are

© PICES 2018. All right reserved.

- Separate the two classes as wide as possible so that there is no misclassification.
- There must be maximum distances of datapoints between the two classes.

The hyperplane is written as

$$y = wx - b \qquad (1)$$

With the value of output whether malware or benign stored in the y variable, also known as classification label [11], w holds the number of datapoints in the plane, x holds the variables and b represents whether it moves it in or out of origin. With proper dataset hyper planes can be described by the following equations:

$$w.x - b = 1 \qquad (2)$$

Equation 2 represents anything on or above this boundary is of one class, with label 1.

$$w.x - b = -1 \qquad (3)$$

Equation 3 represents anything on or below this boundary is of other class, with label -1.

Geometrically, the distance between these two hyperplanes is $2/\|w\|$, so as to maximize the distance between the planes we have to minimize $\|w\|$.

V. FUTURE WORK

The future of malware will grow exponentially. Some of viruses achieve their aims by psychologically manipulating the victim into running unsafe codes or they self-launch themselves without human intervention. To stop the malicious files from reaching the user, machine learning for detection of malware can be used in server rather than using it directly on client side. This would allow the server to analyze the files and documents and provide actions for user's sake. The distinction between servers and client machines has become blurred as servers and workstations have converged to allow multiple processes and connections. Since both, server and workstations will have to run on same versions, they tend to share same vulnerabilities. This could result in major attack into the whole system.

The detection of malware can be done in the Android system [3,5] also, but the mobile devices' do not authorize the users for having malware in the device, which would be easily detectable with their functionality.

Hence, server computing can be used for detection of malware.

VI. CONCLUSIONS

Malware detection on server using distributed machine learning, helps us to precisely predict an unknown file for malware contents. This helps the end-user to safely download its requested file. With the malware growing larger and larger day by day, malware detection mechanisms must grow powerful, with the various machine learning algorithms we will be able to easily detect the file containing malware.

And this approach with server computing [18] enables us to have user-friendly environment in detection of malware.

REFERENCES

[1] Naser Peiravian and Xingquan .Machine Learning for Malware Detection Using Permission and API Calls 2013 IEEE 25th International Conference on Tools with Artificial Intelligence

[2] 2010 International Conference paper on Pattern Recognition Malware Detection on Mobile Devices using Distributed Machine Learning

[3] Dimitrios Damopoulos, Sofia A. Menesidou1, Georgios Kambourakis, Maria Papadaki, Nathan Clarke and Stefanos Gritzalis Evaluation of anomaly-based IDS for mobile device using machine learning classifiers.

[4] Wiley Online Library (wileyonlinelibrary.com).

[5] 8th International Conference on Next Generation Mobile Applications, Services and Technologies, (NGMAST 2014), 10-14 Sept., 2014.

[6] Wen-Chieh Wu Shih-Hao Hung DroidDolphin: a Dynamic Malware Detection Framework Using Big Data and Machine Learning

[7] A. Eliës, "Distributed logic programming for artificial intelligence", AI Communications, vol. 4, no. 1, pp. 11-21, March 2011.

[8] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[9] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining Inference and Prediction, Springer, 2009.

[10] MachineLearning: What it is and why it matters, 09 2016, [online] Available: www.sas.com.

[11] B.-Y. Zhang, J.-P. Yin, J.-B. Hao, D.-X. Zhang, S.-L. Wang, "Using Support Vector Machine to Detect Unknown Computer Viruses", International Journal of Computational Intelligence Research, vol. 2, no. 1, 2014.

[12] Nwokedi Idika, P. Aditya Mathur, A Survey of Malware Detection Techniques, CA: West Lafayette:Department of Computer Science, pp. 3-10, 2015.

[13] https://sucuri.net/website-security-platform/malware-scanning-and-detection

[14] Opendns. [Online]. Available: http://www.opendns.com/

[15] Malware domain list. [Online]. Available: http://malwaredomainlist.com/

[16] P. K. Chan, R. Lippmann, "Machine learning for computer security", Journal of Machine Learning Research, vol. 6, pp. 2669-2672, 2014.

[17] N. Idika, A. P. Mathur, A survey of malware detection techniques, Purdue University, pp. 48, 2007.

[18] Zhang, H. (2013). Architecture of Network and Client-Server model. arXiv preprint arXiv:1307.6665. [2]. Kambalyal, C. (2010). 3-tier architecture. Retrieved On, 2. [3]. Kratky, S., & Reichenberger, C. (2013). Client/Server Development based on the Apple Event Object Model. Atlanta. [4].

© ΠCES 2018. All right reserved.