Predictive Analytics with online data for WSO2 Machine Learner with the support of Ensemble method

Heshani Herath, Ishani Pathinayake, Ashani Diaz and Indujayani Karthigesu, Lakmal Rupasinghe, Krishnadeva.K,
Chethana Liyanapathirana, Sripa Vimukthi Bannakkotuwa
Research Group, Sri Lanka Institute of Information Technology, Sri Lanka
*Contact: lakmalr@gmail.com, phone +94-77-756 1061
heshani7.herath@gmail.com, phone +94-71-739 0012
ishanipathinayake@gmail.com, phone +94-71-074 9024

ABSTRACT—Different types of malware prevail in a world of innumerable computer users who constantly struggle with threats from malware sources such as the internet, local networks, portable drives and so on. Security specialists and malware developers have been at a tug of war with each other as security specialists try to use all possible techniques to overcome the problems arising from malicious software while malware developers try to bypass these implemented security measures. It can be observed from records that each day, nearly 1 million new malware threats are released to the world. Therefore it is evident that there is an indispensable requirement of a proper malware identification mechanism. Typically, antivirus software are based on a signature definition system which keeps updating from the internet and thus keeping track of known malware. While this was sufficient sometime back, it does not cater to the current requirement of identifying malware. Due to the advancement in technology, malware developers have been able to create malware that are able to obfuscate themselves thus giving birth to polymorphic malware. In this study we closely observed the behaviour of malware, tried to understand how they work, their different types, dissemination of malware and detecting/defending mechanism in order to contribute to the process of security enhancement and came up with the solution of integrating Machine Learning to our current problem.

Keywords: Malware family classification, predictive analytics, HTTP Streaming data

I. INTRODUCTION

The Malware is a topic widely spoken in the world of information technology. Malware, being the short term for malicious software consists of code snippets, scripts, active content and etc. The term has invaded the digital world in such a manner that everybody knows its meaning and has taken steps to prevent their computers from being infected by them as they cause severe problems leading to the loss of privacy, increase unauthorized access to system resources and other abusive behaviour and even access to the system unstable. Malware can reach the systems in different ways and through multiple media. Reports show that the most common entrance of malware is through the downloading process via the internet. Malware acts in many different ways. There are some malware that will not totally harm the system but will affect the performance and create overload process, certain other malware can act as spyware in which case the malware will hide itself in the system and the average anti-virus software will not be able to detect them. These hidden malware will send critical information about the computer to the source. By looking at the above challenges it is imperative to carry out a thorough investigation to understand the behaviour of malware for better detection.

II. OBJECTIVES

There have been some efforts to use few machine learning and data mining for the purpose of identifying new or unknown malicious code. But we are mainly focused on identifying malicious code in http data stream. Malicious software in its various forms constitutes a serious threat to Internet security threat. Therefore, this area has received much consideration in the research community, and many different concepts and techniques for the analysis of malware have been proposed. Our goal is, provided that the system predicts a url to be malicious, and classifying the malware families.

In this paper, our goal is to get the lowest number of false positives as possible, by using a simple combination of various stages of the different versions of simple algorithm. Other automated classification algorithms could also be used. We are going to extract data from http streams and going to create feature table with abnormal parameters with the use of extracted data.

A set of characteristics is calculated for the HTTP data streams in training or testing datasets on the basis of many possible ways of analysing a malware. Then classifying the malware using machine learning algorithm. We used three datasets a training dataset, a test dataset. The training and test data must be representative of the web servers to be protected, and the attacks used for testing need to illustrate the range of attacks existing today.

III. SYSTEM OVERVIEW

This section describes the system architecture and the design of the proposed system. It mainly focused on few specific areas. HTTP stream (Online data stream) can contain millions of malwares and important data for day today usage as well. The proposed system mainly focusing on capture those data which comes along with a HTTP stream. System will identify parameters which come along with the HTTP data stream. Those parameters can be changed depend on the stream. Following we have listed some common parameters which comes along with http stream.

Buffer size	The amount of memory allocated for sending content
Client port	A range of client ports for communication.
version	A string specifying the version of HTTP to use
Host name	The host name or IP number for the machine.
port	The socket port number.
Proxy host	The host name of the proxy server
Proxy port	The port number of the proxy server
Stream type	Whether the stream is a text stream, image stream, video etc.

TABLE I. PARAMETERS OF HTTP DATA STREAM

The proposed system will identify parameters of the HTTP data stream and recognize the variations of each parameter. If there is any unusual behaviour system will notify that. Since malwares have polymorphic behaviour there is high possibility even virus guards won't detect the malware. A computer malware is a self-replicating computer program that operates without the consent of the user. It spreads by attaching a copy of itself to some part of a program file, such as a spread sheet or word processor. Malwares also attack boot records and master boot records, which contain the information a computer uses to start up [1]. A polymorphic malware creates an infection in a computer which is known as a polymorphic infection that creates copies of itself, each with different copy to fool a malware detection and users. Variations are usually various forms of encryption or other signatures to make it more difficult for a malware detection program to find and remove the malware from a computer [2]. Sometimes those polymorphic malwares differ one from the other with only one byte of change. Existing virus guards detect this change and it checks the malicious software with their database which contains malware patterns. If they cannot find the new behaviour in the database it identifies it as a new malware and add it to their database. Later they give it as an update. But if we consider what actually happens there, the pervious malware changes a one byte of it and act like a new malware. But it is not a new malware. It is a member of previous malware's family. If we can classify those malwares into families it would be a more effective than a virus guard. For this purpose only we use machine learning.

For malware family classification the proposed system uses Naïve Bayes algorithm[3-8], Support Vector Machine algorithm (SVM)[9-16] and Random Forest classification algorithm[17-19]. These machine learning algorithms support for real time data classifications. Since the system gets three outputs from above three algorithms the proposed system implements an ensemble method which gives the most accurate prediction.

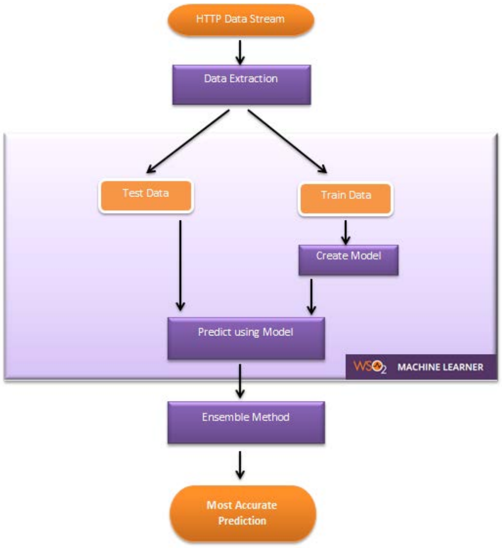


FIGURE II. SYSTEM ARCHITECTURE

IV. METHODOLOGY

This section includes detailed descriptions about the techniques and mechanism employed to make this project a reality. The descriptions include how software implementation of our project is carried out, what are the materials and data needed, and how they will be collected. It also includes time frames and schedules that are required in achieving its objectives. In addition to them, the research areas that we have identified in order to carry out this project are explained rationally.

Creating another machine learner is not the goal of this research. The existing WSO2 machine learner[25] works pretty well. But it does not have the ability of giving an accurate prediction using streaming data (online data). Our main objective is to develop a machine learner which can give accurate results using streaming data. Here we are mainly going to consider about malware family classification in HTTP data stream.

In order to achieve the goals of the proposed project there are 4 major tasks to be completed. Those are,

1. Incremental learning component
2. Predictive model with http data streams
3. Data visualization component
4. Implement Ensemble methods.

Creating the Incremental learning component

In our proposed solution we create a feature table which contains the most common parameters in a HTTP request. We capture the HTTP data stream and essential data to fill the feature table is extracted from the data stream.

Extracted data is cleansed and divided as training and testing dataset to feed the feature table. We input training dataset and train the machine learner to analyse the feature table and to identify most common values for parameters.

The architecture should identify patterns in the history and update the patterns with incoming data without catastrophic forgetting. To accomplish this, the proposed research will create the incremental learning component.

The idea of incremental learning with streaming data focuses on two objectives:

1. Identifying patterns in the recent history.
2. Updating the patterns with incoming data without catastrophic forgetting.

Furthermore the proposed new Machine Learner will automatically detect odd data and remove it from processing. Therefore it will result in high- accuracy and best quality output.

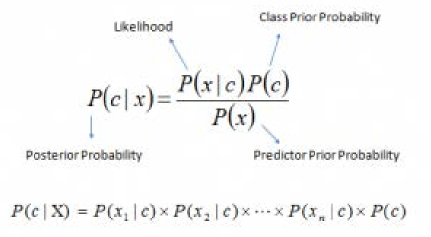
Predictive model with streaming data

Most common algorithms to classify malware families are Naïve Byes [3-8], SVM [9-16] and Random Forest Algorithm [17-19].

Naïve Byes Algorithm

Naive Bayes [3-8] is an eager learning classifier and it is sure fast. It used for making predictions in real time.

It has the multi class prediction feature. Here we can predict the probability of multiple classes of target variable.



$$P(c\mid x)=\frac{P(x\mid c)P(c)}{P(x)}$$

$$P(c\mid X)=P(x_1\mid c)\times P(x_2\mid c)\times\cdots\times P(x_n\mid c)\times P(c)$$

FIGURE II. NAÏVE BAYES ALGORITHM

SVM Algorithm

Support Vector Machine” (SVM) [9-16] is a supervised machine learning algorithm which can be used for both classification or regression challenges.We can plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. By finding the hyper-plane that differentiates the two classes we perform classification.

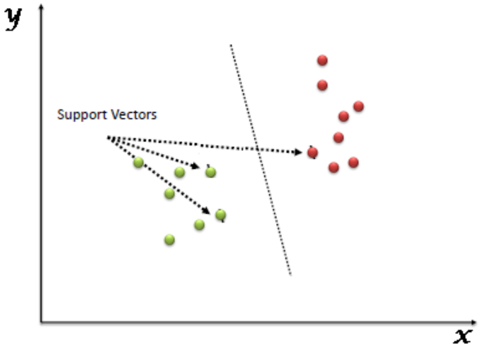


FIGURE IIII. SVM ALGORITHM

Random Forest Algorithm

Random Forest[17-19] is a versatile machine learning method capable of performing both regression and classification tasks. We grow multiple trees as opposed to a single tree in CART model in Random Forest. To classify a new object based on attributes, each tree gives a classification and we say the tree “votes” for that class. The forest chooses the classification having the most votes (over all the trees in the forest) and in case of regression, it takes the average of outputs by different trees.

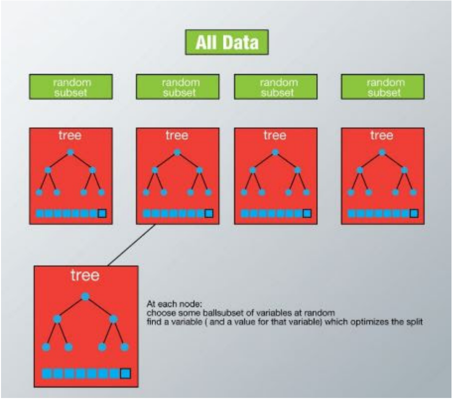

All Data
random subset
random subset
random subset
random subset
tree
tree
tree
tree
tree
At each node:
choose some ballsubset of variables at random
find a variable (and a value for that variable) which optimizes the split

FIGURE IV. RANDOM FOREST ALGORITHM

Current WSO2 machine leaner is capable of classifying data using Naïve Byes [3-8], SVM[9-16] and Random Forest [17-19]Algorithms. Once the machine learner is trained testing dataset is used to classify the malware family. By using testing dataset we identify whether there are any abnormal behaviours in the feature table. If we identify any abnormalities in feature table via our proposed machine learner we detect the malware family and predict the malware family classification. Malware family is classified using above 3 algorithms.

According to the results of this predictive model, the performance of the machine learner will be increased with online streaming data for malware family classification.

Data visualization component

Once the data is processed it will be appeared to the user in a format which any user can understand. Data will be presented in a way anyone can easily understand with a basic computer knowledge and basic English knowledge. Proposed system use multiple visualizations to explore the data such as scatter plots, histograms, Trellis charts, cluster diagrams and so on.

Implement Ensemble methods

The whole idea is to employ multiple apprentice methods and combine their predictions. The aim set of methods is the combination of several predictions based estimators constructed as a learning algorithm in order to improve the generalization / robustness on one estimator. It provides better predictive accuracy than a single learning algorithm. Final prediction will be done by taking the weighted vote of the predictions of the gathered algorithms [20-24].

Naïve Bayes algorithm [3-8], Random Forest algorithm [17-19] and Support Vector Machine (SVM) algorithm[9-16] exist in WSO2 Machine Learner is used to get the Malware family classification of the HTTP data stream.

After data is processed through these algorithms and an outcome is predicted, Ensemble methods[20-24] for Machine Learning is implemented to combine these multiple algorithms such that a better predictive accuracy could be achieved than what could be achieved from a single learning algorithm[10]. Final prediction is done by taking a weighted vote of the predictions of the combined algorithms.

V. Future Work

By improving the proposed research we are planning to identify the malware rather than identifying the malware family. It will be a huge task and the team is expecting to focus on some specific malware in order to identify them. Furthermore by improving this research work we can use to bio medical science as well. We can improve this system in order to identify changes in DNA and genes.

VI. Conclusions

The Hypertext Transfer Protocol (HTTP) has become a universal transport protocol. Timely and accurate detection of anomalies in massive HTTP data streams plays a major role in preventing machine failures, intrusion detection, and dynamic load balancing. Advanced malwares are posing a severe threat to the internet and computer systems. We considered the application of techniques from machine learning, data mining to the problem of detecting and classifying unknown malicious on http data streams. For the malware classification we are use three algorithms such as Naïve Bayes algorithm [3-8], Random Forest algorithm [17-19] and Support Vector Machine (SVM) algorithm[9-16]. Finally Ensemble method[20-24]implemented to obtain accurate prediction. We hope that such a strategy for detecting and classifying malicious will improve the security. We can improve these results in the future.

References

[1] [Online]Available: https://www.symantec.com/avcenter/reference/striker.pdf [Accessed: March 8th, 2016]

[2] [Online]Available: http://www.computerhope.com/jargon/p/polyviru.htm [Accessed: March 8th, 2016]

[3] [Online]Available: https://en.wikipedia.org/wiki/Naive_Bayes_classifier [Accessed: March 8th, 2016]

[4] [Online]Available: http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab4-NaiveBayes.pdf [Accessed: March 8th, 2016]

[5] [Online]. Available: Constrained Clustering, Sugato Basu, Ian Davidson, Kiri L. Wagstaff [Accessed: March 8th, 2016]

[6] [Online]Available: http://users.ics.aalto.fi/juha/papers/TRUSTCOM15_Android.pdf [Accessed: March 8th, 2016]

[7] [Online]Available: http://www.aicit.org/JCIT/ppl/JCIT%20VOL7NO5_part25.pdf [Accessed: March 8th, 2016]

[8] [Online]Available: http://www.analyticsvidhya.com/blog/2015/09/naive-bayes-explained/ [Accessed: March 8th, 2016]

[9] [Online].Available: Support Vector Machine, Jason Weston [Accessed: March 8th , 2016]

[10] [Online].Available:Making Large-Scale SVM Learning Practical,Thorsten Joachims[Accessed: March 1st, 2016]

[11] [Online].Available: Support Vector Machines and Metamorphic Malware Detection,Thorsten Joachims[Accessed: March 1st, 2016]

[12] [Online]Available: http://www.svm-tutorial.com [Accessed: March 8th, 2016]

[13] [Online]Available: http://www.iis.sinica.edu.tw/page/jise/2015/201505_11.pdf [Accessed: March 8th, 2016]

[14] [Online]Available: http://phdthesis.uaic.ro/PhDThesis/Cimpoe%C8%99u,%20Mihai,%20%20Classification%20Algorithms%20for%20Malware%20Detection.pdf [Accessed: March 8th, 2016]

[15] [Online]Available: http://www.ijcsit.com/docs/Volume%206/vol6issue04/ijcsit2015060424.pdf [Accessed: March 8th, 2016]

[16] [Online]Available: http://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=1405&context=etd_projects [Accessed: March 8th, 2016]

[17] [Online]Available: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm [Accessed: March 8th, 2016]

[18] [Online]Available: https://arxiv.org/ftp/arxiv/papers/1205/1205.3062.pdf[Accessed: March 8th, 2016]

[19] [Online]Available: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm [Accessed: March 8th, 2016]

[20] [Online].Available:https://www.toptal.com/machine learning/ensemblemethods-machine-learning [Accessed: March 3rd, 2016]

[21] [Online].Available: http://in.mathworks.com/help/stats/ensemblemethods.html?requestedDomain=www.mathworks.com [Accessed: March 1st, 2016]

[22] [Online]. Available: https://datajobs.com/data-science-repo/Ensemble-Methods[Lior-Rokach].pdf [Accessed: March 8th, 2016]

[23] [Online]. Available: Ensemble Learning, Thomas G. Dietterich [Accessed: March 8th, 2016]

[24] [Online].Available: http://web.engr.oregonstate.edu/~tgd/publications/mcsensembles.pdf [Accessed:March 8th, 2016].

[25] [Online].Available:
https://docs.wso2.com/display/ML100/Introducing+Machine+Learner
[Accessed:March 8th, 2016].