International Journal of Scientific Research & Engineering Trends
Volume 3, Issue 5, Sept.-2017, ISSN (Online): 2395-566X

Malicious Web URL Classification using Evolutionary Algorithm

Anshika Bansal
Scholar CSE, AITR, Bhopal
Email: anshika23bansal@gmail.com

Abstract **– The proposed works provide an evolutionary algorithm based machine learning approach for malicious URL classification. This work use BAT algorithm for features extraction forms all URLs from the URL list. It includes identifying and determining the features based approach to classify different types of malicious URLs. The proposed approach employed the extracted data to Support Vector Machine classifier for classification of testing data to labels as benign, spam and phishing.**

Keywords **– BAT, URL, Spam, Phishing.**

I. INTRODUCTION

A "malicious web page" refers to a web page that contains malicious content that can exploit a client–side computer system. Malicious website may be used as a weapon by cybercriminal to exploit various security threats such as phishing, drive-by-download and spamming. Malicious Web sites are hurdle on the way ofInternet security. And used as a weapon to mount various security threat like phishing, drive-by-download and spamming. To handle there is need to develop an automatic system to recognized malicious website. To derive detection models for malicious web pages, distinguishing features of benign and malicious web pages are analyzed. Nevertheless, these artifacts are under constant evolution. This evolution is typically because of two reasons. First one, cyber-criminals constantly revamp their strategies to craft attack payloads in malicious web pages not only aimed at making attacks more complex but also to evade existing countermeasures. Second, benign web pages evolve because of new content, new functionalities, or changes to the underlying technologies used in building the web pages.

Proposed framework initially use combine Web page data set contain Alexa white listed URL, phishing URL provided by Phish Tank and spam url. In proposed framework initially initialized data set with random partition. Then BAT approach tends to generate random feature for malicious url. This BAT is recertifying through randomness. If randomness of relevant feature is low then it's acceptable otherwise feature is not to be consider. Acceptable Feature is denoted as relevant feature and apply for classify malicious. If malicious URL have high false negative rate then whole process is initialed by random partition and if malicious URL have low true positive rate then new feature is generated on same partition.

This dissertation conclude the proposed technique for malicious URLclassification that is based on one of the evolutionary algorithm known as BAT and apply SVM for classification. With the help of BAT we calculate the degree of uncertainty and again on the basis of randomness with the classify the Web URL, web url having higher entropy, is regarded as the "malicious", and generate the alarm.

II. WEB APPLICATION

Web Applications are software applications deployed by the World Wide Web. They use a single client-server model, and run in a Web browser on the client computer. Once a new release of a Web Application is installed on the server, this release is available to all users. This immediate deployment characteristic is probably one of the most powerful characteristics of a Web Application. There are different names in use for what here is called a Web Applications. Names in use are Web Sites, Web-based applications and Web Applications. Some authors are also using different names to indicate different types of Web Applications. In this article the term Web Application is used to represent all types.

III. WEB USAGE DATA

The use of online data held in the first visiting team patterns recorded on a website. You can also include visions, titles, customer cakes, and modify customer data and inquiries Any response to other customers do, while on the site. They are easy to control data compilation and comfort into three sections, this means that the network host records, records of hosting the door and the client browser web server important information about the network that uses dig those records in the public access to sites maintained by many users. Each of the records containing the user's IP address, time Pettion, Uniform Resource Locator, the HTTP status code, etc. Of course, the information has been collected in several standard formats such as log file format, and how the calculation log files, etc.

Gateway server, well known as a web proxy server acts as a gateway server for users and servers. To reduce the time and the agent receives the web page entry is a useful tool for customers who visit these websites often, and with the reception and the agent is also useful to have a

complete picture of the traffic load in the server and the client. Proxy Server has the ability to view the complete text applications using hypertext transfer of different users to different Web servers protocol. Use these proxy surf activities for a group of similar customers, and determine who are involved are the same proxy server and thoughtful analysis. Agent is available on the client side is useful to collect usage information for the user on the client side. You can also see this factor as a web browser and has the ability to identify the tasks performed by customers. These logs to collect information from a particular client, many sites. In the customer information capture critical information regarding these records on the Internet or a door, for example, to return to mouse clicks or page using the spacebar back also used.

- **Preprocessing**

This method is used to manage records were actual weaving operation before the actual mining was the main intention of this method is equipped with a full recognition or activities Web sessions. By storing the complete information to the web server for the behavior of all Web server logs, client access. Said the staff show time: The client 2.3.During this total process and IP address does not match the question of the role of known properties. Most Web logs are a group of state and an access point and a client or in chronological order according to the increased demand. This method applies to all log files to ensure that information about the courses on the Internet. The roads are included in the information processing and cleaning, and to identify the stage of customer recognition.

- **Data cleaning**

It is including the step of implementing each fishing information on the Internet that are not suitable for drilling intentions For example, records: Search for Gholam pilot rate subject (such as JPG and GIF) any other files that can be incorporated into applications VARLET network. Or even browse robots and spiders Internet session. When the application and graphics comfortable buying online Pension plans eliminate spider and robot navigation should be and must be a clear example. This habit and practice, for example, have been transferred to the device from a distance, quoting an agent or to ensure access to the robots.txt file. However, some robots and very fake agent sent from the client in the HTTP request. In this case, the heuristic based on browsing behavior can be used to split sessions literal client Android courses and this is reflected in the first gestures of the navigation screen qualified search engine in a tree to symbolize the structure and not intended reference site. (Originator site provides the user after returning to reports). Disclosure of the proposed heuristic based on the assumption that the above categories of sailors.

Web logs recorded during user interactions can not be extracted directly. Therefore, it is necessary to deal with HTML documents in case of arrival. File type consists of images and records files in standard Internet resources. The image in any of these formats can be a file, such as BMP GIF, JPG or. special code includes hypertext transfer cases that are useful to represent the availability or unavailability of the desired item indicates. Getting rid of the events that have 200-299 Minors Act fruitful, and the rest when using Web logs. Any other formats such as HTML URLs, ASP, JSP, etc ... are removed from the records.

- **Client Recognition:**

From the perspective of describing the behavior of users to determine whether the first user therefore be treated as anonymous as mentioned above. One way to determine the user's client is the IP address. Therefore, applications for the same IP can be treated as the same user. Additional information about customers that can help better understand the behavior patterns of users seeking. Access to the Internet for many users using the same IP factor is the same, but the type of agent may be different. Therefore, we can not accept that the function of each type of agentive symbolizes the same IP address of the client.

- **Session Recognition**

It is understood that the client has passed the site more than once every time the session length is farsighted. The purpose of the meeting is a recognition of the separation of the soul records of user network sessions for their arrival. Regarded as a new special session if the difference between the time a request of two contiguous registers the user is greater than the threshold limit. In this work, we have set the default to 30 minutes time limit threshold value.

- **Other Preprocessing Tasks**

Processing tasks used depends on the intention of mining. a full path is used to find the physical access path between web pages. It can be verified from the field reference in the Web logs to see from any page of the application has arrived. If the reference is not available, and the site link structure can also help evaluate the path of users. The objective of establishing the transaction is to create a web page visited significant required for each user group. Thus, the function of the transaction is the recognition of the increased separation according to a number of smaller operations or a combination of smaller to larger transactions. Some methods have been proposed to recognize the offer for a distance of, for example signal, the maximum signal forward and time window

IV. PROPOSED METHODOLOGY

The proposed works provide an evolutionary algorithm based machine learning approach for malicious URL classification. This work use BAT algorithm for features extraction forms all URLs from the URL list. It includes identifying and determining the features based approach to classify different types of malicious URLs. The proposed approach employed the extracted data to Support Vector Machine classifier for classification of testing data to labels as benign, spam and phishing.

International Journal of Scientific Research & Engineering Trends
Volume 3, Issue 5, Sept.-2017, ISSN (Online): 2395-566X

V. COMPARATIVE ANALYSIS

It is the measure the accuracy of the training model to classify the test data set. Accuracy Analysis can be measured by four value i.e. TP, TN, FP and FN directly or indirectly. These four value define:

i) True Positive (TP) = If a URL entry is proven present in a class and the proposed model also classifies that URL in that class, the result is considered true positive.
ii) True Negative (TN) = If a URL entry is proven absent in a class and the proposed model also proves the absence of that URL in that class, the result is considered true negative.
iii) False Positive (FP) = If the proposed model indicates the presence of a URL in a class who actually does not contains that URL, the result is considered false positive.
iv) False Negative (FN) = If the proposed model indicates the absence of a URL in a class, who actually belongs to that class, is considered false negative.

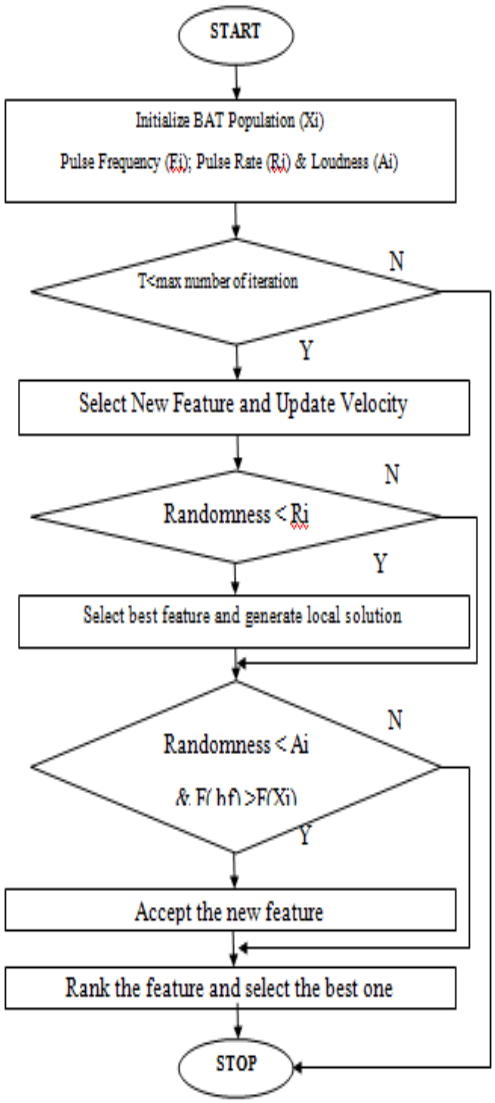

Fig.1. Flow Graph

Comparing to our algorithm (BAT) and PSO. The results of our algorithm's classification rate performance are shown in Table 1, and the results of PSO-SVM classification performance are shown in second column Table 1. We can find our algorithm outperforms than PSO. And we can see our algorithm overcomes some problems existing in PSO.

Table 1: Comparing To BAT Algorithm and PSO

	BAT	PSO
TPR	92.23	78.93
FNR	5.76	3.76
Accuracy	90.47	80.21

VI. COMPARATIVE GRAPH

The below figure is a comparative graph for the both methods with three parameters. the conclusion comes from this graph is the BAT gives the batter results from the PSO. Here the FNR is equal but other parameter shows the result is batter then PSO.

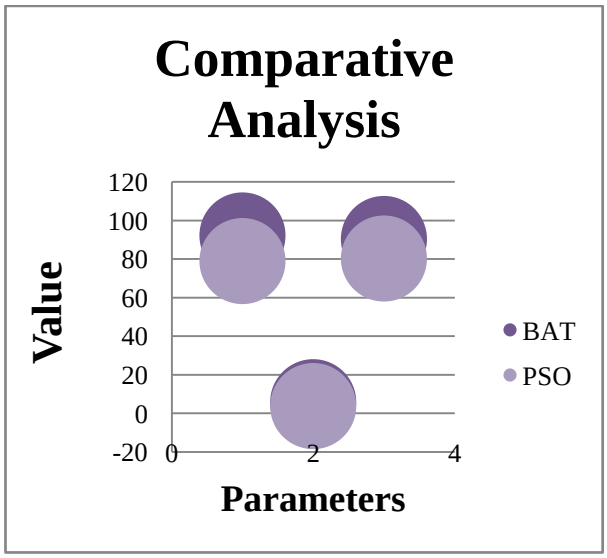

Fig.2. Comparative Graph

Accuracy is defined as the percentage of correctly classified result. In this dissertation accuracy is defined by the correctly classification of the URLs in their respective class i.e. Benign, spam and phishing. In terms of true positive and true negative accuracy is defined as:

$$\text{Accuracy} = \frac{\text{Total Positive}}{\text{Total Assessments}} = \frac{TP + TN}{(P + N)}$$

Where
TP = Number of instances that are correctly predicted to their actual class.
TN = Number of instances that correctly rejected.
P+N = Total number of instances to be classified.

VII. CONCLUSION

Detection of malicious web has become a necessary and hot topic of research as numbers of internet users are

increasing at a high pace. There are lots of challenges regarding this detection process. First the number of online URL is very large. Second web environment uses diverse platform and difficult to find security solution for them. Third now threats are become more and more complex and used various obfuscation techniques to bypass detection techniques. The existing detection techniques are focused only on single type of attacks only. New generated malicious web pages exploit multiple types of attacks for targeting the client. Cloaking type of attacks is difficult to detect because these web respond differently to browser and crawler. Size of web is a big challenge in the process.

REFERENCES

[1] M. Wu and M. Yang, "Privacy Preservation for Detecting Malicious Web Sites from Suspicious URLs," *2011 International Conference on Business Computing and Global Informatization*, Shanghai, 2011, pp. 400-403.

[2] Y. Fukushima, Y. Hori and K. Sakurai, "Proactive Blacklisting for Malicious Web Sites by Reputation Evaluation Based on Domain and IP Address Registration,"*2011IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications*, Changsha, 2011, pp. 352-361.

[3] D. Kent and L. M. Liebrock, "Statistical detection of malicious web sites through time proximity to existing detection events," Resilient Control Systems (ISRCS), 2013 6th International Symposium on, San Francisco, CA, 2013, pp. 192-197.

[4] L. Vu, P. Nguyen and D. Turaga, "Firstfilter: A cost-sensitive approach to malicious URL detection in large-scale enterprise networks," in*IBM Journal of Research and Development*, vol. 60, no. 4, pp. 4:1-4:10, July-Aug. 2016.

[5] S. B. Rathod and T. M. Pattewar, "A comparative performance evaluation of content based spam and malicious URL detection in E-mail,"*2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS)*, Bhubaneswar, 2015, pp. 49-54.

[6] Z. Li-xiong*et al.*, "Malicious URL prediction based on community detection,"*Cyber Security of Smart Cities, Industrial Control System and Communications (SSIC), 2015 International Conference on*, Shanghai, 2015, pp. 1-7.

[7] M. S. Lin, C. Y. Chiu, Y. J. Lee and H. K. Pao, "Malicious URL filtering A big data application," *Big Data, 2013 IEEE International Conference on*, Silicon Valley, CA, 2013, pp. 589-596.

[8] H. K. Pao, Y. L. Chou and Y. J. Lee, "Malicious URL Detection Based on Kolmogorov Complexity Estimation," *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on*, Macau, 2012, pp. 380-387.

[9] M. K. K. Leung, A. Delong, B. Alipanahi and B. J. Frey, "Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets," in*Proceedings of the IEEE*, vol. 104, no. 1, pp. 176-197, Jan. 2016.

[10] M. Nickel, K. Murphy, V. Tresp and E. Gabrilovich, "A Review of Relational Machine Learning for Knowledge Graphs," inProceedings of the IEEE, vol. 104, no. 1, pp. 11-33, Jan. 2016.

[11] S. Gunduz, B. Arslan and M. Demirci, "A Review of Machine Learning Solutions to Denial-of-Services Attacks in Wireless Sensor Networks," 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), Miami, FL, 2015, pp. 150-155.

[12] N. M. De Mel, H. H. Hettiarachchi, W. P. D. Madusanka, G. L. Malaka, A. S. Perera and U. Kohomban, "Machine learning approach to recognize subject based sentiment values of reviews,"2016 Moratuwa Engineering Research Conference (MERCon), Moratuwa, 2016, pp. 6-11.

[13] "Spam URLs," [Online]. Available: http://www.joewein.de/sw/bl-text.htm. [Accessed 10 September 2016].