ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

Heuristic Based Approach for Fraud Detection using Machine Learning

Nida Khan[1], Manliv Kaur[1], Riddhi Panchal[1], Prashant Kumar Rai[1], Nilesh Rathod[2]

B.E, Department of Information Technology, Rajiv Gandhi Institute of Technology, Mumbai, India[1]

Professor, Department of Information Technology, Rajiv Gandhi Institute of Technology, Mumbai, India[2]

ABSTRACT: Internet has become a useful part of our regular day to day life as we do almost all of our social and financial activities online. Today, everyone is heavily reliable on internet and online activities such as online shopping, online Banking, online booking, online recharge and many more. Phishing is a type of social engineering attack that targets a user sensitive information through a phony website that appears similar to a legitimate site, or by sending a phishing email .Heuristic based approach is to produce a solution in a reasonable time that is good enough for solving the problem .Heuristic approach defines that it may produce results by themselves, or they may be used in conjunction with optimization algorithms to improve their efficiency (e.g., they may be used to generate good seed values). With the limitation in the existing system we are introducing additional features through the heuristic approach which is simpler and effective than the earlier approaches. This is mainly used for real-world applications and one of this is used in fraud detection on an online platform. Internet environment and diversification of available web services, web attacks have increased in quantity and advanced in quality. Heuristics approach through machine learning underlie the whole field of Artificial Intelligence and the computer simulation of thinking, as they may be used in situations where there are no known algorithm. The heuristic-based detection technique analyses and extracts phishing site features and detects phishing sites using that information. It is imperative to detect and act on such threats in a timely manner. However, blacklists cannot be exhaustive, and lack the ability to detect newly generated malicious URLs. To improve the generality of malicious URL detectors, machine learning techniques have been explored with increasing attention in recent years.

KEYWORDS:: heuristic approach, phishing sites, blacklist detection

I. INTRODUCTION

The advent of new communication technologies has had tremendous impact in the growth and promotion of businesses spanning across many applications including online-banking, ecommerce, and in social networking. Unfortunately, the technological advancements come coupled with new sophisticated techniques to attack and scam users. The most common method to detect malicious URLs deployed by many antivirus groups is the black-list method. Blacklists are essentially a database of URLs that have been confirmed to be malicious in the past. This database is compiled over time. The limitations of traditional security management technologies are becoming more and more serious given this exponential growth of new security threats, rapid changes of new IT technologies, and significant shortage of security professionals. The first module is the URL and DNS matching module which contains a whitelist, which is used to increase the running time and decrease the false negative rate. Our white-list maintains two parameters, domain name and corresponding IP address. Whenever a user accesses a website, then the system matches the domain name of the current website with white-list. If the domain of the current website is matched with the white-list, then the system matches the IP address to take the decision. When the user access a website which is already present in the white-list, then our system matches the IP address of the corresponding domain to check the DNS poisoning attack. Our white-list starts with zero; it means that at the beginning, there is no domain in the list and the white-list starts increasing once a user accesses the new webpages. If the user is accessing the website for the first time, then the domain of the website will not be present in the white-list. In that case, our second module starts working. The second module is the phishing identification module, which checks whether a webpage is phishing. this then extract the hyperlinks from the webpage and apply our phishing detection algorithm . Our phishing detection algorithm examines the features from the

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

hyperlinks to take the decision. After checking the legitimacy, if the website is phishing, then the system shows the warning to the user. Moreover, if the website is legitimate, then the system updates the domain in the white-list.

II. RELATED WORK

There are multiple previous approaches to detect phishing attacks. Some of those approaches include Google Safe Browsing, PhishNet, Phish Guard, Spoof Guard, Baitalarm and BlockLayout Similarity etc.

- Google Safe Browsing: This approach uses the blacklist urls to discover the phishing attack . A sample url is taken as input and checked in the blacklist repository. If the url is present in the blacklist repository, the url is termed as suspicious URL, else it is a legitimate website.
- PhishNet: This approach overcomes the problems related with the blacklists. It has two major steps such as generation of URL variations relative to the original ones which grows the blacklist as well as a data structure which assigns each score to URL based on similarity with existing URLs .
- PhishGuard: This research implements an algorithm ObURL to rate the suspicious web sites based on the visual appearance of the web pages. This algorithm identifies White List Test, Black List Test, IP Address Test, Shorten URL Test, DNS Test, Pattern Matching Test.

Recent years have witnessed innovative applications of machine learning in cyber security. For example, present a survey on the usage of machine learning and data mining techniques for Cyber Security intrusion detection.For example an empirical analysis of different machine learning techniques for Malicious URL Detection in 2007, at a time when neither features nor machine learning models for this task had been extensively explored. gave a broad overview of Phishing and its problems, but do not extensively survey the feature representation or the learning algorithms aspect. focused on primarily feature selection for Malicious URL Detection. Malicious URL Detection is closely related to other areas of research such as Spam Detection. This conducted a comprehensive survey in 2012, wherein they identified different types of spam(Content Spam, Link Spam, Cloaking and Redirection, and Click Spam), and the techniques used to counter them. They categorized the Spam Detection techniques into Content based Spam Detection (using lexical features such as Bag of Words and Natural Language Processing techniques), Linkbased spam detection (utilizing the information regarding the connectivity of different URLs) and other miscellaneous techniques. Spam Detection is heavily reliant on processing the text in an email and utilizing natural language processing for analysis. These techniques are not directly useful for Malicious URL Detection, unless they are used to draw inference about the context in which the URL has appeared. Despite some overlap between the techniques used for spam detection and malicious URL detection, spam detection techniques largely qualify as techniques that use context-based features for detecting malicious URLs. Other recent learning based spam detection surveys include many of which focus on spam appearing in online reviews.

Disadvantage:

- Attackers use many other simple techniques to evade blacklists including: fast-flux, in which proxies are automatically generated to host the web-page; algorithmic generation of new URLs.
- Attackers can often simultaneously launch more than one attack, which alters the attack-signature, making it undetectable by tools that focus on specific signatures.
- Attackers will also try to obfuscate the code so as to prevent signature based tools from detecting them.

III. PROPOSED ALGORITHM

ID3 ALGORITHM:

Iterative Dichotomiser that is ID3 is a decision tree learning algorithm which was invented by Ross Quinlan which is used for generation of decision tree from datasets.ID3 is the precursor to the C4.5 algorithm and is typically used in fields like machine learning and natural language processing domains.

The ID3 algorithm consists of original set S as the root node.On each iteration of algorithm,it iterates through every unused attribute of the set S and calculates the entropy H(S) or the information gain IG(S) of that attribute.The attribute with the smallest entropy value or largest information gain value is selected.The set split S is then split by the selected

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

attribute to produce subsets of the data.For example age is less than 50, age is between 50 and 100, age is greater than 100.The algorithm continues recursion on each subset, considering only attributes that are never slected before.
Recursion on a subset may stop in one of these cases:

- every element in the subset belongs to the same class (+ or -), then the node is turned into a leaf and labelled with the class of the examples
- there are no more attributes to be selected, but the examples still do not belong to the same class (some are + and some are -), then the node is turned into a leaf and labelled with the most common class of the examples in the subset
- there are no examples in the subset, this happens when no example in the parent set was found to be matching a specific value of the selected attribute, for example if there was no example with age >= 100. Then a leaf is created, and labelled with the most common class of the examples in the parent set.

Throughout the algorithm, the decision tree is constructed with each non-terminal node representing the selected attribute on which the data was split, and terminal nodes representing the class label of the final subset of this branch.
Summary of Algorithm is as follows:

1. Using data set S calculate the entropy of every attribute.
2. Split the set S into subsets using the attribute for which the resulting entropy (after splitting) is minimum (or, equivalently, information gain is maximum)
3. Make a decision tree node containing that attribute
4. Recurse on subsets using remaining attributes.

IV. PSEUDOCODE

```
ID3 (Examples, Target_Attribute, Attributes)
    Create a root node for the tree
    If all examples are positive, Return the single-node tree Root, with label = +.
    If all examples are negative, Return the single-node tree Root, with label = -.
    If number of predicting attributes is empty, then Return the single node tree Root,
    with label = most common value of the target attribute in the examples.
    Otherwise Begin
        A ← The Attribute that best classifies examples.
        Decision Tree attribute for Root = A.
        For each possible value, v_i, of A,
            Add a new tree branch below Root, corresponding to the test A = v_i.
            Let Examples(v_i) be the subset of examples that have the value v_i for A
            If Examples(v_i) is empty
                Then below this new branch add a leaf node with label = most common target value in the examples
            Else below this new branch add the subtree ID3 (Examples(v_i), Target_Attribute, Attributes – {A})
    End
    Return Root
```

V. CONCLUSION AND FUTURE WORK

In this project, we proposed a URL based phishing attack technique that employs URL-based features .We have added new features by analyzing the websites which are phising websites along with URL based features that were used in the previous studies. We have generated classifiers using machine learning algorithms and found that ID3/DECISION TREE Algorithm are good classifiers.The technique which we have proposed in our project can help naïve users to detect the phising sites based on the features and also help in providing them with security for personal information and reduce damage caused by phising sites and phisingattacks.It can detect new and temporary phishing sites that evade existing phishing detection techniques, such as the blacklist-based technique.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

REFERENCES

1. R. K. Nepali and Y. Wang, "You look suspicious!!: Leveraging visible attributes to classify malicious short urls on twitter," in 2016 49th Hawaii International C]
2. C. Seifert, I. Welch, and P. Komisarczuk, "Identification of malicious web pages with static heuristics," in Telecommunication Networks and Applications Conference, 2008. ATNAC 2008. Australasian. IEEE, 2008, pp. 91–96onference on System Sciences (HICSS). IEEE, 2016, pp. 2648–2655.
3. Huang, Huajun, Junshan Tan, and Lingxi Liu. "Countermeasure techniques for deceptive phishing attack." New Trends in Information and Service Science, 2009. NISS'09. International Conference on. IEEE, 2009
4. PhishTank, [Online] Available:http://www.phishtank.com
5. DMOZ, [Online] Available: http://rdf.dmoz.org/rdf/
6. Anti Phishing Working Group. (2015. March.) APWG PhishingActivity Trend Report 2nd Quarter 2014. [Online]. Available: http://docs.apwg.org/reports/apwg_report_q2_2010.pdf
   Huang, Huajun, Junshan Tan, and Lingxi Liu. "Countermeasuretechniques for deceptive phishing attack." New Trends in Information and Service Science, 2009. NISS'09. International Conference on.IEEE, 2009.
   Ma, Justin, et al. "Beyond blacklists: learning to detect malicious websites from suspicious URLs." Proceedings of the 15th ACM SIGKDDinternational conference on Knowledge discovery and data mining.ACM, 2009
   Nguyen, Luong Anh Tuan, et al. "A novel approach for phishingdetection using URL-based heuristic." Computing, Management and Telecommunications (ComManTel), 2014 International Conference on. IEEE, 2014.
   Wikipedia. (2015. March) Uniform Resource Loactor. Avaliable:http://en.wikipedia.org/wiki/Uniform_resource_locator Kausar, Firdous, et al. "Hybrid Client Side Phishing Websites
7. Detection Approach." International Journal of Advanced ComputerScience and Applications (IJACSA) 5.7 (2014).
   Sunil, A. Naga Venkata, and Anjali Sardana. "A pagerankbaseddetection technique for phishing web sites." Computers & Informatics (ISCI), 2012 IEEE Symposium on. IEEE, 2012.
   Mohammad, Rami M., FadiThabtah, and Lee McCluskey."Intelligent rule-based phishing websites classification." Information Security, IET 8.3 (2014): 153-160.
   Canali, Davide, et al. "Prophiler: a fast filter for the large-scaledetection of malicious web pages." Proceedings of the 20th international conference on World wide web. ACM, 2011.
8. Xiang, Guang, et al. "Cantina+: A feature-rich machine learning framework for detecting phishing web sites." ACM Transactions onInformation and System Security (TISSEC) 14.2 (2011): 21.
9. WANG, Wei-Hong, et al. "A Static Malicious Javascript Detection Using SVM." strings. Vol. 40. 2013.
10. L. Ladha and T. Deepa, "Feature selection methods and algorithms,"International journal on computer science and engineering, vol 3, no5, 2011.
11. Hou, Yung-Tsung, et al. "Malicious web content detection by machine learning." Expert Systems with Applications 37.1 (2010): 55-60.
12. Cao, Ye, Weili Han, and Yueran Le. "Anti-phishing based on automated individual white-list." Proceedings of the 4th ACM workshop on Digital identity management. ACM, 2008.