Efficient Malicious URL based on Feature Classification

Samridhi Sharma[1], Shabnam Parveen[2]

1. M.Tech Student, 2Assistant Professor, Computer Science & Engineering Department
JMIT, Radaur/Kurukshetra University, India

1. samridhisharma91@gmail.com

2 .er.shabnam786@gmail.com

9416333410

Abstract— Deceitful and malicious web sites pretense significant danger to desktop security, integrity and privacy. Malicious web pages that use drive-by download attacks or social engineering techniques to install unwanted software on a user's computer have become the main opportunity for the proliferation of malicious code. Detection of malicious URL has become difficult because of the phishing campaigns and the efforts to avoid blacklists. To look for malicious URL, the first step is usually to gather URLs that are live on the Internet. Then different algorithms are applied to detect malicious URL. This paper is about classifying URL based on features using machine learning techniques OneR, ZeroR, and Random Forest.

Keywords—. Machine Learning, Feature Extraction, Benign, Malicious, Web Pages, Classification Module, Attacks.

INTRODUCTION

The internet has become the medium of option for public to search for information, conduct business, and enjoy entertainment. At the same time, the internet turns out to be the most important stage used by miscreants to attack users. The most commonly used example is drive by download attack. In this attack, attackers insert different modes of attack in the web pages to which malicious URLs direct and once the victim clicks on a malicious URL, they are taken to that web page without notice. Then the attacker may steal any of the victim's information that is saved on the host computer, which may lead to grave financial loss. When malicious URLs are sent by friends, victims are more likely to click them. In addition to drive-by-download exploits, attackers also use social engineering to trick victims into installing or running un trusted software. As an example, consider a webpage that asks users to install a fake video player that is presumably necessary to show a video (when, in fact, it is a malware binary). Another example includes fake anti-virus programs. These programs are expanded by web pages that alert users into thinking that their machine is infected with malware, alluring them to download and execute an actual piece of malware as a remedy to the claimed infection. The web is growing rapidly and is a very large place, in which new pages (both begnin and malicious) are added at formidable place.

There has been lot of changes and phases in the history of malicious software since it has been exposed and detected in hosts and networks, preliminary from virus which is a self-Replicating adware but not self-transporting moving to worm, which is a self-replicating and self- transporting and going more for other. The figure of malware attack is increasing sharply with the rapid increase in complexity and interconnection of rising information systems. When the user clicks on the URL it is most likely to become a target. To prevent users from visiting such URL much may be malicious or contain illegal content, large amount of research generated by the security industry is done.

According to one study by the Gartner Group [McCall 2007], damage caused by the phishing in the United States is $3.2 billion loss in 2007, amid 3.6 million victims lessening for the attacks, a enormous raise from the 2.3 million the year previous to. Moore et al [Moore and Clayton 2007] provided details that the loss suffered by the consumers and businesses in 2007 in the US unaccompanied was about $2 billion.

A major percentage of those losses were basis by one mainly infamous group, called as the "rock phish gang" that uses toolkits to create a large number of unique phishing URLs, putting more pressure on the correctness and precision of blacklist-based anti-phishing techniques. New, previously unseen malicious executables, polymorphic malicious executables using encryption and metamorphic malicious executables adopting obfuscation techniques are more complex and difficult to detect. At present, most commonly used malware detection software make use of signature-based method and the heuristic based method to identify threats.

Signatures are strings of bytes which are short and exclusive to the programs. There use is to recognize scrupulous threats in executable files, records of boot, or memory. The disadvantage, this signature based method is not effective next to customized and unidentified malicious executables this is due to the signature extraction and generation process. Heuristic-based method is more complex than signature based detection techniques, the disadvantage of this method is that time consuming and still fails to detect new malicious executables.

The main outcomes of malicious content can be broadly grouped into the following three categories:

_ Phishing

_ Deceptive advertising

_ Computer infection for unauthorized use

A. Phishing

Phishing is an attack whereby an attacker tries to gain user's personal information by trying to trap the user into entering identifying and account information for a legitimate service. This method primarily targets financial and payment service sectors. Phishing attacks focusing on the economic and payment service sectors account for about 71% of the phishing attacks during that time, as indicated by the statistics.

B. Deceptive advertising

With this method of attack users are prompted to buy counter feit goods at low cost. The main example for this attack includes email spam advertising which will often show a pricelist as well as a URL to the web page where the goods may be purchased.

C. Computer infection for unauthorized use

The use of Botnets is done to exploits the machines by propagating when the user installs software. This can be done via many attack vectors, including that of drive-by downloads. When the user clicks on the link then the attacker exploits the user browser by installing the unwanted software in the background without the knowledge of user.

RELATED WORK

The presented work done on the detection of malicious URL can be broadly classified into three categories, specifically the blacklist based methods, the text based methods, and the URL based methods.

The *blacklist based methods* there are numerous blacklists existing which are a collection of malicious URLs and can be queried before visiting a page. The examples of backlists include Phish tank, Google Safe browsing, McAfee Site Advisor, Fortinet, URL lookup tool and WebsenseThreatSeeker Network. In order to have maintained blacklist variety of techniques are available such as honeypots, user feedbacks and crawlers. The advantage of blacklist method is it is accurate and simple and the disadvantage includes firstly the results produced are slow because of the direct verification process secondly it does not guarantee that every new Malicious URL will be in the backlist as it requires users to discover malicious URL.

The *text based methods* examine the text of the matching web page of a URL to detect whether the URL is malicious. The text provided by the Web pages for detection is very useful and of much consequence. For example, Provos et al. [1] discovered malicious URLs using features from the content of the equivalent URLs, such as the presence of definite javascript and whether iFrames are inappropriate. Moshchuk et al. [2] used the antispyware tools to examine downloaded trojan executables to identify malicious URLs. Byung-ik kim Km et al. [3] analyzes JavaScript density, frequency, entire JavaScript entropy, and entropy of each characteristic for the detection of malicious websites. The advantage of content based methods is when one has to perform an offline detection and analysis; they are not competent of online detection. The disadvantage in online detection is that they often incur major latency, because examining and evaluate page text repeatedly costs much computation time and resource.

Most recently, the *URL based methods* use only the URL structures in detection URL Patterns are of much importance as with these patterns users can find the malicious actions. URL based features are now days used repeatedly to detect the malicious URL. McGrath and Gupta [4] they noticed disparity among normal URLs and phishing URLs in some features, such as the length of the URL, domain name length ,number of dots in URLs. With these features a classifier is build for phishing URL detection. Yadav et al. [5] analyzed additional features, which include the dissimilarity in bi-gram distribution of domain names between normal URLs and malicious ones. Their results conclude that normal URLs and malicious ones indeed have noticeable dissimilarity in the features extracted from URLs themselves alone. Le et al. [6] demonstrate that by using only the URL lexical features can maintain most of the performance in phishing URL detection. Kan and Thi[7] also worked on the lexical features of the URL only, but their work tries to group to normal URLs, such as news, business, and sports, instead of detecting malicious URLs from normal ones. Egan, S.et al, [8] concluded that the lightweight classifier is only slightly lower than that of fully-featured classification. They can define the normal behavior through static analysis of the browser behavior and compare with the browser behavior visiting a malicious web page, then determine whether a web page is malicious or not. Jian Cao et al., [9] have focused on forwarding based features along with URL and graph based features in order to train a detection model. They assess the arrangement employing concerning 100,000 early memos amassed from SinaWeibo, which is the biggest OSN website in China. Their study concludes that the forwarding base features are more effective than conventional features because of the high accuracy and low false positive rate HodaEldardiry et al.[10] has proposed a malicious insiders detection prototype which includes two types of activities blend-in anomalywhere malicious insiders try to behave similar to a group they do not belong. For this behavioral inconsistencies across these domains are observed which include logon, device, file, http, email sent and email received, and unusual change anomaly where malicious insiders exhibit changes in theirbehavior. Fusion algorithm is used to combine anomaly from multiple source of information

Hence this section presents the work done in the previous for the detection of malicious URL. The conclusion is most of the work done in this area in based on the URL features hence we will focus on the same.

CLASSIFICATION METHODS

This section briefly describes the various classification methods. As there are many classification algorithms but here we will describe only few of them. In machine learning there are two types supervised learning and unsupervised learning. Supervised learning presume that training examples are classified (labeled by class labels)Unsupervised learning focus on the examination of unclassified examples. In both cases, the objective is to build a *model* for the entire dataset, or to discover one or more patterns that hold for some part of the dataset.

Navie Bayes

It is generally used in spam filters. This method has been used form the ages in the information retrieval and text classification due to its probabilistic nature. The concept used in it is explanation the previous probability of each class, *P*(*Ci*),and the conditional probability of each feature value known the class, *P*(*aj*j*Ci*). It calculates approximately these measures by counting in training dataset the frequency of occurrence of the target class and of the feature values for all target class. Then, it uses the Baye's rule to calculate the latter probability of each target class; an unidentified instance is given which is returning prediction of the target class with the maximum such value:

C = argmax *Ci P*(*Ci*)□*jP*(*a j*j*Ci*).

Support Vector Machine (SVM): SVMs are widely considered as recent models for binary classification of high dimensional data. They are trained to increase the edge of correct classification, and the resultant decision boundaries are vigorous to slight perturbations of the feature set, thus as long as a hedge beside over fitting. The better generalization capabilities of SVMs have been stand out by both theoretical studies and experimental successes. The decision rule in SVMs is given in terms of a kernel function K(x, x ') that calculate the resemblance between two feature vectors and non-negative coefficients $\{\alpha_i\}_{i=1}^{n}$ that point to which training examples lie close to the decision boundary. SVMs classify new examples by calculating their distance to the decision boundary. Up to a constant, this distance is given by:

$$H(x) = \sum_{i=0}^{n} a_i(2y_{i-1})K(x_{i\,,}x),$$

where the sum is over all training examples. The sign of this distance indicates the side of the decision boundary on which the example lies. In practice, the value of h(x) is threshold to predict a binary label for the feature vector x. SVMs are trained by first identifying a kernel function K(x, x ‘) and then computing the coefficients α_i that increase the margin of accurate classification on the training set. The necessary optimization can be devised as an example of quadratic programming, a predicament for which many efficient solvers have been developed.

Logistic Regression:

This is a straightforward parametric method for binary classification where examples are classified based on their distance from a hyper plane decision boundary. The decision rule is expressed in terms of the sigmoid function $\sigma(z) = [13]^{-1}$ which converts these distances into probabilities that feature vectors have positive or negative labels. The conditional probability that feature vector x has a positive label y= 1 is the following:

$P(y=1/x)=\sigma(w.x+b)$

EXPERIMENTS

1). DATA SET:The data set used in this research consisted of known malicious and known benign URLs. For malicious dataset, we obtained the data from Squidblacklist[11], a web service which provides lists of malicious URLS. Benign dataset was obtained from Alexa,[12] which is a web service that ranks web sites based on traffic generated. For example, web sites such Google and Facebook will be ranked higher in their list of websites compared those that are less frequented. Higher traffic generated web sites are less likely to be malicious because such sites are well preserved due to their fame among internet. The collected data was grouped into training with 460 instances and 534 attributes and testing data set. Training set is defined as a set of data used to discover potentially predictive relationships. The training set is used at the initial stage of the proposal to determine patterns or similarities between the different set of data obtained. The testing set is used to verify the set of patterns or similarities that was exposed during the training stage. The output of testing includes 220 malicious URL and 240 begnin URL.

2). FEATURE SELECTION

A. Host-based features: Host based features are those that require the use of exterior sources. The sources of information used in these features are WHOIS data, IP address information, and Domain Name properties which are briefly described as:

WHOIS data - It is the protocol through which query is made and responses are given by the database. It includes details when the domain is registered; expiration date registrars. With all this information available classifier can conclude how new the domain is and whether or not the domain belongs to an individual already associated with other malicious URLs.

IP address information - Each device in the computer network is assigned an IP address. It includes two features `host or network interface identification and location addressing. It is used to verify whether or not an IP address is in a blacklist. It includes the feature like hosting of the website and includes the IP address prefix and the AS number. This allows a specific ISP’s IP prefix to be flagged as malicious by the classifier. In addition to this it is associated geo location of the IP address.

Domain Name - It represents an Internet Protocol resource which includes host computer to access the Internet. Today the number of active domains reached 271 million.

B. Lexical features: These are the features of the URL that refer to the actual text of a URL and contain no outside information. These features are helpful as malicious URLs often “look” different than benign ones to experts. Features that fit in to this set consist of statistical information concerning lengths of features, tokens, numbers of delimiters and directory structure. This information is valuable as it is obfuscation resistant. To get the numerical statistic two terms are used firstly term frequency is the number of times a word appear in document divided by total no of words in that document and secondly inverse document frequency defined as the logarithmic of total number of documents divided by the number of documents with term t in it.

C. Conclusion

Host-based features, lexical features or a combination of both, are then run through a classifier which will then present a forecast as to whether the URL is malign or benign. We represent this as 0 for malign and 1 as begnin.

3).Classifiers

In this section we have make the use of three supervised algorithms which are ZEROR, ONER, and Random Forest. As the URLs are collected they are then fragmented by the use of string to word vector convertor. After that we train the classifier and then test it .

A. ZERO R

ZeroR is the simplest association method that relies on the target and overlooks all its features. ZeroR classifier plainly predicts the popular group (class).Although there is no predictability manipulation in ZeroR, it is functional for delineating a baseline presentation as a benchmark for supplementary association methods.

B. ONER

OneR, also known as "One Rule", is a simple algorithm, yet accurate, This classification algorithm generates one rule for each predictor in the data set, and it then selects a rule with smallest total error as the "one rule". To create a rule for a predictor, The algorithm has to construct a frequency table for each predictor against the target class.

- Find Count of each value of target(class) in the dataset
- Calculate the most frequent class
- Make the rule allocate that class to this value of the predictors
- Find the total error of the rules for each predictor in the data
- Select the predictor with the smallest total error.
- Find the best predictor which possesses the smallest total error using OneR algorithm.

C. RANDOM FOREST

Random forests are an ensemble discovering method for association, regression and supplementary tasks, that work by constructing a multitude of decision trees at training period and outputting the class that is the mode of the classes (classification) or mean forecast (regression) of the individual trees. Random forests correct for decision trees' custom of over fitting to their training set.

The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set $X = x_1, \ldots, x_n$with responses $Y = y_1, \ldots, y_n$, X_i Represents the input attributes, input URLs and Y represents the class {Malicious, Benign},

1. Split each URL according to their Latent features.
2. For each feature calculate frequency in each URL as Term Frequency
3. Create Feature Matrix $X_i^{''}$by executing Inverse document Frequency for all urls
4. Classify Feature Matrix $X_i^{''}$using Random Forest by
5. For $b = 1, \ldots, B$:
6. Sample, with replacement, n training examples from X, Y; call these $X_i^{''}$ Y_b.
7. Train a decision or regression tree f_b on $X_i^{''}$, Y_b.
8. After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x':

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(x')$$

or by taking the majority vote in the case of decision trees.

RESULTS AND ANALYSIS

In our examinations, the classifier was trained and tested on comparable number of benign URLs as malicious URLs. The ratios of benign-to-malicious URLs do not differ considerably in training and testing. Such conditions can arise after the classifier is used in a disparate manner than it was trained. For example, presume that a established spam filter is utilized to remove URLs from dubious emails alongside product advertisements. With such URLs by now flagged, the aim of the classifier should shift to noticing phishing locations that do not continue in nearly the alike abundance as locations that merely vend spam-advertised products. The table below shows the result obtained for the classifying the URL for algorithms ZeroR, OneR, Random Forest.

Table1: Percentage of Correct and Incorrect URL

Classifier	Correct	Incorrect
ZERO R	52.17391304347826%	47.82608695652174%
ONER	54.56521739130435%	45.43478260869565%
RANDOM FOREST	91.52173913043478%	8.478260869565217%

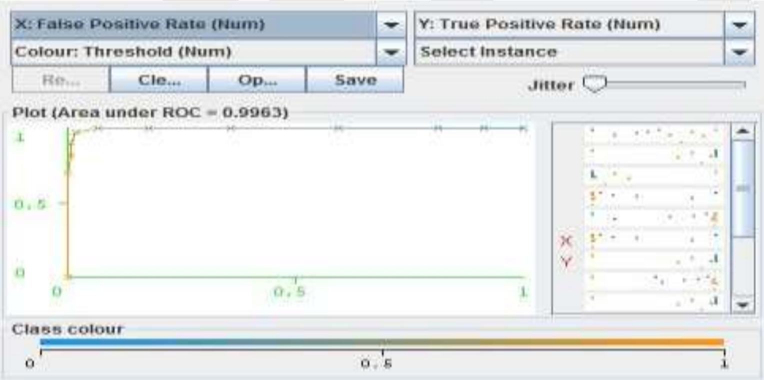
X: False Positive Rate (Num)
Y: True Positive Rate (Num)
Colour: Threshold (Num)
Select Instance
Plot (Area under ROC = 0.9963)

Fig1: Classification Results of Random Forest based Classifier showing ROC

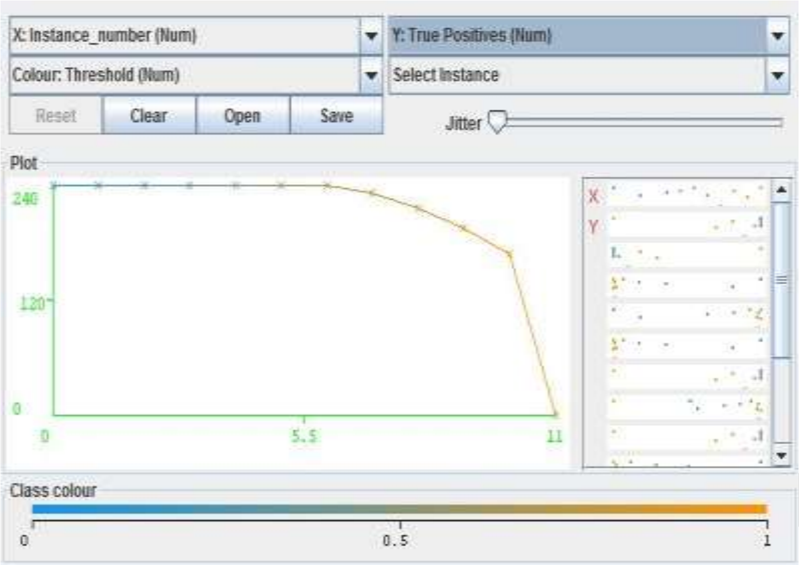
X: Instance_number (Num)
Y: True Positives (Num)
Colour: Threshold (Num)

Fig 2: Classification Results of Random Forest based Classifier showing True Positive per instance

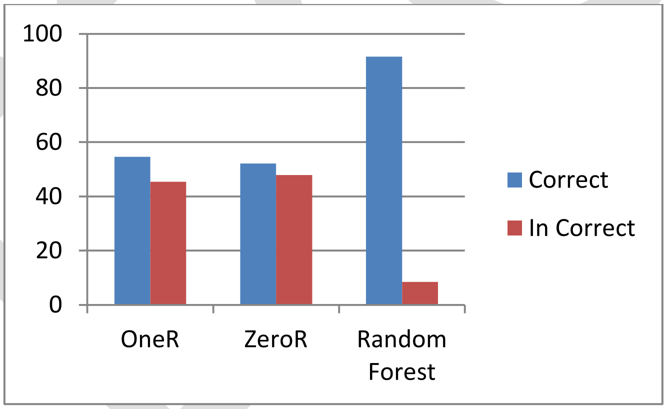

Fig 3: Graph showing the percentage for OneR, ZeroR and Random Forest**.**

CONCLUSION

This work presents the evaluation of three supervised contraption discovering models namely ZeroR, OneR and Random Forest for Malicious URL association, to notice URL as whichever malicious or not. All the supervised methods were trained and requested to a colossal number of URL's and manually tear into two classes. In a nutshell, Random forest based method display reassuring presentation alongside classification above 91% and Zero R and One R being 52% 54% respectively. As the number of URLs increased the accuracy rates for every single kind of contraption discovering ideal can be improved.In upcoming we will incorporate bit.ly like URL shortening in dataset for enhancing the finished association complexity. Enhancing presentation of Classifier can additionally be one more span of research. We will tolerate to enhance the scrutiny for discovering harmful malicious traffic and clustering clients infected by malware such as malicious bots, worms, virus and tolerating to vanquish weaknesses. We will additionally work on leading examination consolidated alongside a main IDS or malicious detection arrangements to assess the presentation of the method

REFERENCES:

[1]. Provos, N., Mavrommatis, P., Rajab, M.A., Monrose, F.: All your iframes point to us. In: Proceedings of the 17th Conference on Security Symposium, pp. 1–15. USENIX Association, Berkeley, CA, USA (2008)

[2]. Moshchuk, A., Bragin, T., Gribble, S.D., Levy, H.M.: A crawler-based study of spyware in the web. In: Proceedings of the Network and Distributed System Security Symposium (NDSS'06).The Internet Society, San Diego, California, USA (2006)

[3]. Byung-ik kim et. al Suspicious malicious website detection with strength analysis of a javascript obfuscation International journal of advanced Science and technology 2011

[4]. McGrath, D.K., Gupta, M.: Behind phishing: an examination of phisher modi operandi. In: Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats,pp. 4:1–4:8. USENIX Association, Berkeley, CA, USA (2008)

[5]. Yadav, S., Reddy, A.K.K., Reddy, A.N., Ranjan, S.: Detecting algorithmically generated malicious domain names. In: Proceedings of the 10th Annual Conference on Internet Measurement.IMC '10, pp. 48–61. ACM, New York, NY, USA (2010)

[6]. Le, A., Markopoulou, A., Faloutsos, M.: Phishdef: url names say it all. In: Proceedings of the 30thIEEE International Conference on Computer Communications, Joint Conference of the IEEEComputer and Communications Societies, pp. 191–195. IEEE, Shanghai, China (2011)

[7]. Kan, M.-Y., Thi, H.O.N.: Fast webpage classification using url features. In: Proceedings of the14th ACM International Conference on Information and Knowledge Management. CIKM '05,pp. 325–326.
ACM, New York, NY, USA (2005)

[8]. Egan, S. et al, in "An evaluation of lightweight classification methods for identifying malicious URLs" 2011

[9]. Jian Cao, Qiang Li, YuedeJi, Yukun He, and Dong Guo. *"Detection of Forwarding-Based Malicious URLs in Online Social* Networks." International Journal of Parallel Programming (2014): 1-18.

[10]. HodaEldardiry, Evgeniy Bart, Juan Liu, John Hanley, Bob Price, and Oliver Brdiczka. "Multi-domain informationfusion for insider threat detection." In Security and Privacy Workshops (SPW), 2013 IEEE, pp. 45-51. IEEE, 2013.

[11]. To get the mailicious URL http://www.squidblacklist.org/

[12]. To get the begnin URLhttp://www.alexa.com/topsites/global;1