Classification of benign and malicious websites by selecting important URL features using LASSO

Tsai, Tzung-Lin[†1] Goutam Chakraborty[†2]

Abstract: On the Internet, users visit unknown web-sites, of which some are malicious, which implant malwares into users' computers through drive-by-downloads technology. With increased use of internet browsing, web security is an important issue and consequently an important research topic. The motivation of this study is to classify malicious web-sites from benign ones from their URL features, for automatic filtering of website. URL data, which has many characteristics, were collected. Some characteristics, but not all, are relevant for the task to classify the site as malicious or benign. We first converted textual features into numeric data so that the numeric values truly represent the original feature. We then selected the relevant features for our classification task, by using least absolute shrinkage and selection operator (LASSO). Finally, the data is used to train a support vector machine (SVM) classifier. A ten-fold validation is used to estimate the performance. We did achieve a high precision of classification, with more than 90% correct classification.

Keywords: Web security, Malicious Web Sites, features selection, LASSO, classification.

1. Introduction

In recent years, number of websites is growing exponentially because of a massive growth in application areas and services, such as social networking, blogs, and e-commerce [9]. With the advancements of technology, high-speed internet connection available and wireless hot-spots everywhere, World Wide Web has become a platform to support a wide range of internet criminal activities too such as spam-advertising, financial fraud, and malware implanting [8].

Although the criminal motivation behind these acts are different, the common aim is to attract unsuspecting users to the website which can be visited via links through e-mail, web search, or links from other websites. All of these links require the user to click on the so-called Uniform Resource Locator (URL). Thus, each time users decide to click on an unfamiliar URL, they must pay attention to the website's address and using their intuition and/or experience evaluate the associated risk that might be encountered. With novice users the chance of getting connected to a malicious URL is high.

To avoid these dangerous events to occur, many kinds of security technologies have been developed. The most common and popular technology is to construct a blacklist, to protect the browsers from the phishing websites by marking those malicious websites as "dangerous". Blacklist has a basic access control mechanism to limit an user to access websites whose information, such as IP address, URL, and domain, are in the blacklist. Blacklists are distributed to other users who use the technology to block malicious or phishing websites. Recently, blacklisting can be done by online application plug-in such as Google Blacklist and BrightCloud or by Antivirus software such as AVAST and McAfee. However, many malicious websites are not in the blacklist because they are created too recently or never evaluated. Yoshiro et. al. [11] pointed out that traditional blacklist cannot filter the unknown malicious websites which are not contained in the blacklist.

In this work, we decide a website as malicious or benign based on its URL features. There are many features of which some are relevant for the classification task, and other are not. We need to focus on selecting significant features. Otherwise, irrelevant features which would act as noise, would reduce classification accuracy, if included in the classification task. In the beginning, we consider all possible features. From URL features, we adopted lexical features as well as host-based features, which characterize an URL.

The rest of the paper is organized as follows. Section II surveys the related work. Section III explain the proposed classification algorithm. The detail methodology to classify the malicious websites consists of three steps: (1) This includes listing all features we adopt, and how to collect the datasets; (2) Selecting relevant features using feature selection method LASSO; and finally (3) Training of SVM classifier. Section IV presents the experimental results. Here, we use ten-fold classifier to evaluate precision of classification. Section V and Section VI present concluding remarks and future work.

2. Related work

Drive-by-download attack [2, 11, 13] is, by which, malicious websites inject malware onto users' computers when users visit these websites which look like benign/legal websites. Provos et. al. identified four major types of the attack: advertising, third-party widget, web security application, and user contributed content [15]. The flow of drive-by-download attack is depicted in Figure 1. In drive-by-download attack, first, when a user browses the landing sites, he will be directed to a drive-by-download server, usually called hot point. The hot point will identify the vulnerabilities of the user's system and select the weakest one to launch an attack. The attack will command the browser to download malware from the distribution site. Finally, the malware is installed and it executes automatically without user noticing it. Because of the attack

†1.2 Graduate school of Software & Information Science, Iwate Prefectural University, Takizawa city, Japan.

grows rapidly, most researches develop detection (before it is activated) to prevent the attacks.

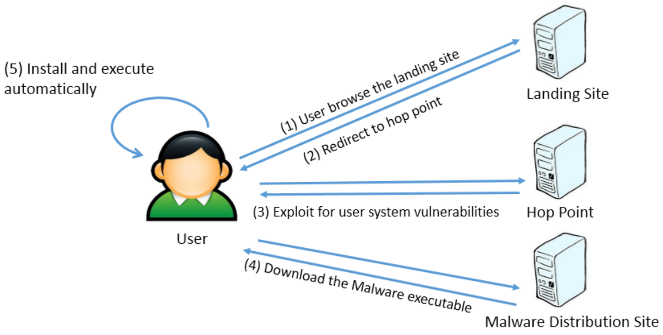

(5) Install and execute automatically
(1) User browse the landing site
(2) Redirect to hop point
Landing Site
User
(3) Exploit for user system vulnerabilities
Hop Point
(4) Download the Malware executable
Malware Distribution Site

Figure 1 Steps how Drive-by-download attack takes place

Blacklisting is a popular and widely used technology. Google blacklists approximately 9500 to 10000 websites per day [18]. However, though blacklisting prevents lots of malicious attacks, it is not effective to protect when the attacking websites are unknown. Blacklisting can be combined with other technologies for better security. We propose to use machine learning to classify malicious websites in real-time while accessing an unknown website [11].

Ma et al. [8] used four datasets and validated the possibility of identifying malicious websites by using three machine learning models: Naïve Bayes, Support Vector Machine with an RBF kernel and regularized logistic regression. Kazemian and Ahmed [6] compared several machine learning models including three supervised classifiers: K-Nearest Neighbor, SVM, and Naïve Bayes; and three unsupervised techniques: Mini Batch K-Means, Affinity Propagation and K-Means. While evaluating detection performance, supervised techniques show 85-97 % classification accuracy. The result of unsupervised machine learning classifiers yielded 0.88 to 0.96 silhouette coefficient. Moreover, Darling et. al. [12] developed a classification systems based on lexical analysis. They collected their datasets by configuring their crawler to collect from six sources and used 87 features for their decision tree based system. However, the main disadvantage was the fact that they used enormous number of features to achieve their results.

The aim of this paper is to use feature selection to select the significant features based on lexical features and host-based features and evaluate the classification ability by using only the selected significant features.

3. Methodology

This section describes the data set we use for the experiment. We collected the data from the following sources: Clean mx [14] for the malicious websites which offer the manually verified malicious URL. We used Open Directory Project (DMOZ) [4] for the benign websites which contain user submitted URL and is the largest directory of the Web. Our input data have 46 thousand unique URLs which include 35 thousand benign URLs and 11 thousand malicious URLs.

The complete experiment framework is shown in Figure 2. In step 2 we collect the website's URLs from Clean mx and DMOZ. In step 3 the features for the training dataset are calculated through various sources. In step 4, the significant features are selected by least absolute shrinkage and selection operator (LASSO). In step 5, we use Support vector machine (SVM) to train the model using only features selected by LASSO. Finally, we classify the malicious and benign websites. The performance is evaluated by accuracy of classification from the average of 10-fold validation.

The following Subsection explains the feature groups and the basis for their selection, how the lasso select the significant features and how the SVM classifer is trained.

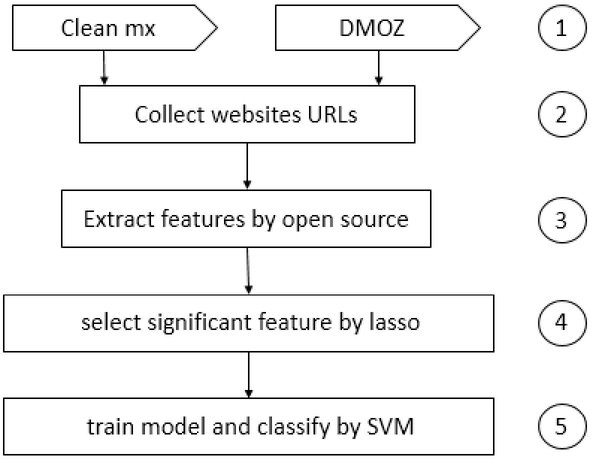

Clean mx
DMOZ
1
Collect websites URLs
2
Extract features by open source
3
select significant feature by lasso
4
train model and classify by SVM
5

Figure 2 The experiment framework.

3.1 Features

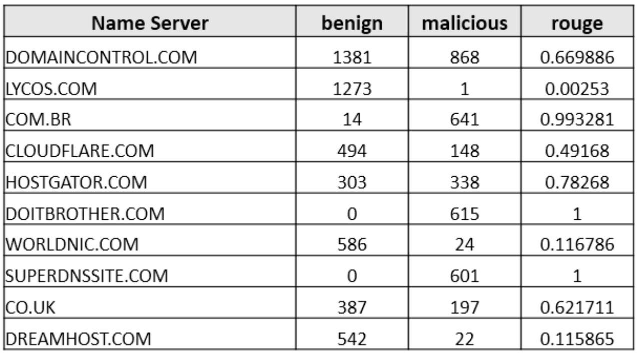
Features used to categorize URLs are either lexical or host-based.

Lexical feature are textual parts of URLs which allow the user to see the differences by reading. The URL has three main parts: the protocol, hostname, and path. For example, let us see the following URL: "http://www.iwate-pu.ac.jp/information/". The protocol is "https://", the hostname is "www.iwate-pu.ac.jp", and the path is "information/". Lexical features are the properties of the URL itself and do not include content of the web page [11]. We use these features for classification purpose. It is suggested to be important by the study of McGrath et al. [3], Ma et al. [8], and Choi et al. [5]. These features include length of each part, average length of hostname token and path token, longest length of hostname token and path token, and Top level domain (TLD) type. In addition, lexical features include dash counts in the hostname, symbol counts in the path, and the numeric ratio of hostname and path.

Host-based features are used to see the hosting of the website and the reputation of the hosting center. The properties of hosts include IP address properties, geographic properties, domain name properties, and WHOIS properties. Our host-based features include domain country, Autonomous system number (ASN), country of ASN, name sever amount, rouge of Name sever, rouge of source, domain create date, and domain update date. The name sever record is relative to domain register. Many domain name registration agents offer very cheap service. Thus, hackers may buy lots of domain name at low cost and perform the malicious operations. As illustrated in Table 1, the name sever is used by more than 500 websites. We can see some name servers are usually used by malicious websites. We can calculate the rouge index of name sever by the following formula

$$Rouge = \frac{\frac{Malicious\ amount}{Total\ Malicious}}{\frac{Malicious\ amount}{Total\ Malicious} + \frac{Benign\ amount}{Total\ Benign}}$$

By the rouge index of a name sever, we judge the probability whether it is a rouge. If the rouge index is close to 1, the name sever is possibly used by malicious websites, and if the rouge index is close to 0, the name sever is possibly operated by a benign domain registration agent.



Name Server	benign	malicious	rouge
DOMAINCONTROL.COM	1381	868	0.669886
LYCOS.COM	1273	1	0.00253
COM.BR	14	641	0.993281
CLOUDFLARE.COM	494	148	0.49168
HOSTGATOR.COM	303	338	0.78268
DOITBROTHER.COM	0	615	1
WORLDNIC.COM	586	24	0.116786
SUPERDNSSITE.COM	0	601	1
CO.UK	387	197	0.621711
DREAMHOST.COM	542	22	0.115865

Table 1 Rouge index of name severs

3.2 LASSO

Lasso is least absolute shrinkage and selection operator. It was introduced by Robert Tibshirani in 1996 based on Leo Breiman's Nonnegative Garrote [10, 16]. Among existing feature selection algorithms, LASSO has been demonstrated as the most practical one because of its robustness and high precision [7, 20]. As it uses linear regression as model, it is very fast. The number of selected features could be tuned giving it flexibility of use. It is known that variable selection and parameter estimation via LASSO are more stable than other subset selection procedure and produce better prediction accuracy compared to other methods of similar efficiency. For these reasons, LASSO is one of the most popular feature selection algorithm among existing ones. It can automatically select the number of variables by shrinking the coefficient values of variables and setting some equal to zero. Besides, LASSO coefficient values determine the importance of different factors, the more its absolute value is, the more important the factor is.

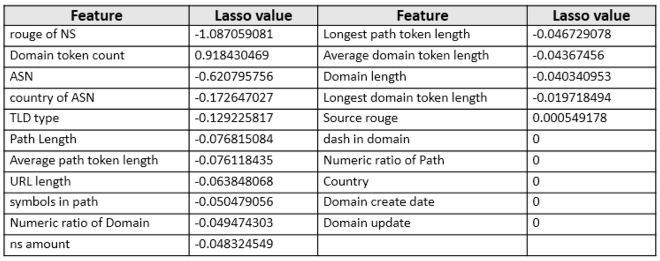
3.3 Support Vector Machines

Support Vector Machines (SVMs), proposed by C. Cortes and V. Vapnik [1], are supervised learning models with associated learning algorithms. In comparison to artificial neural network models, like multilayer perceptions, it is widely used for its better generalization performance. It is regarded as the most effective model for binary classification of high dimensional data. SVMs are trained to find out the maximum-margin hyperplane and the trained classification boundary is robust. Better generalization and classification ability of SVM have been proved by many theoretical works and successful experiments with real data.

4. Results

In the section, we select the significant features using LASSO and evaluate the classification accuracy using SVM classifier. SVM is trained using the selected significant features. The classification result is evaluated using different performance evaluation parameters such as accuracy, precision, recall, and F – Measure. In the experiment, 10-fold cross validation is applied to the data set. In 10-fold cross validation [17], the data set is divided into 10 random partitions. One tenth is used for testing the classifier and nine tenth is used for training the classifier. The training and testing data are changed and classification accuracy is calculated for 10 different sets of testing data. The average value is given.

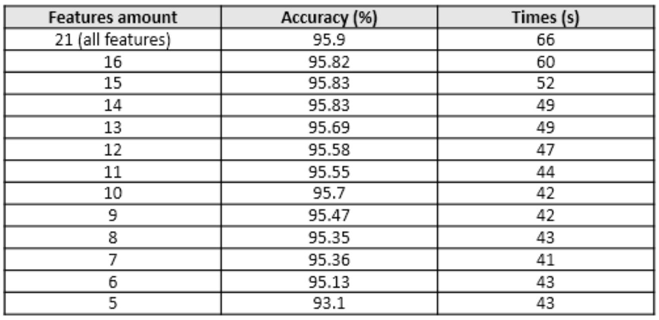
Table 2 shows the result of significant features selected by lasso. We sort the features in descending order without considering the positive and negative. In the result, we choose the 16 significant features in the total features.



Feature	Lasso value	Feature	Lasso value
rouge of NS	-1.087059081	Longest path token length	-0.046729078
Domain token count	0.918430469	Average domain token length	-0.04367456
ASN	-0.620795756	Domain length	-0.040340953
country of ASN	-0.172647027	Longest domain token length	-0.019718494
TLD type	-0.129225817	Source rouge	0.000549178
Path Length	-0.076815084	dash in domain	0
Average path token length	-0.076118435	Numeric ratio of Path	0
URL length	-0.063848068	Country	0
symbols in path	-0.050479056	Domain create date	0
Numeric ratio of Domain	-0.049474303	Domain update	0
ns amount	-0.048324549		

Table 2 Result of Lasso – features and corresponding coefficient values (absolute) in descending order

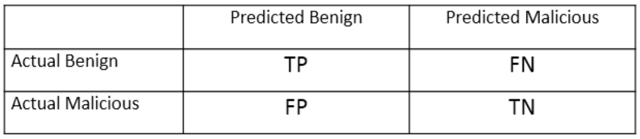
Based on the significant features, we use SVM to evaluate the performance. We choose different numbers of features, from 5 to 16, by descending order of the value of the feature coefficient and evaluate the accuracy of the classifier and computational cost. Table 3 show the relationship of the number of features, accuracy, computational cost (computation time). We can see that when the top 10 significant features are used we can get a good classification accuracy at a low computation cost.



Features amount	Accuracy (%)	Times (s)
21 (all features)	95.9	66
16	95.82	60
15	95.83	52
14	95.83	49
13	95.69	49
12	95.58	47
11	95.55	44
10	95.7	42
9	95.47	42
8	95.35	43
7	95.36	41
6	95.13	43
5	93.1	43

Table 3 The relationship between the number of features and classification accuracy

To illustrate measurement index, we use table 4 to demonstrate precision, recall, F1-measure and accuracy.



	Predicted Benign	Predicted Malicious
Actual Benign	TP	FN
Actual Malicious	FP	TN

Table 4 Binary classification

In the table 4, the meanings of notations TP, FP, FN, and TN are given as follows:

(1) TP (True positive): Actual benign websites classified into “Benign”

(2) FN (False Negative): Actual benign websites classified into “Malicious”

(3) FP (False Positive): Actual malicious websites classified into “Benign”

(4) TN (True Negative): Actual malicious websites classified into “Malicious”

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{TP + FP + FN + TN} \quad (1)$$

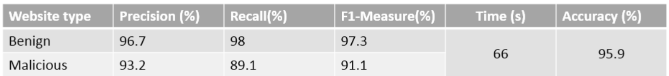
$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

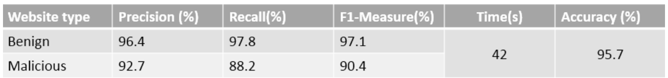
The accuracy has been defined in equation (1). F1 is a weighted index when both Precision and Recall are considered with equal weightage. Precision, Recall, F1 are defined in equations (2) ~ (4).

Table 5 shows the classification result based on all features we considered in the experiment. Table 6 shows the classification using the top 10 significant features. According to the result, the accuracy in both the cases are above 95% in detecting malicious websites. However, the classifier should be able to classify both, malicious and benign web-sites of the data set. Ability of the classification model to classify both malicious and benign websites are presented through precision, recall, and F1-measure. In our result, the F1-measure of the model for benign websites are both above 97% and for malicious websites are above 90 %. Besides, the result of top 10 significant features selected by LASSO takes less computation time and could yield similar result as when the whole feature set is used.



Website type	Precision (%)	Recall(%)	F1-Measure(%)	Time (s)	Accuracy (%)
Benign	96.7	98	97.3	66	95.9
Malicious	93.2	89.1	91.1		

Table 5 The classfication result of all the 21 features



Website type	Precision (%)	Recall(%)	F1-Measure(%)	Time(s)	Accuracy (%)
Benign	96.4	97.8	97.1	42	95.7
Malicious	92.7	88.2	90.4		

Table 6 The classification result using top 10 features

5. Conclusions

The paper presents a website classification based on URL and host-based features. LASSO is used for feature selection. We experiment with various features and select the significant features to classify the websites. The data is collected from the web source, which is the real-world data. The experimental results show that the classification accuracy is higher than 95 %. By the proposed feature selection approach the computation time is reduced and we could achieve similar classification accuracy.

6. Future work

In our experimental result, the classification accuracy using the top 10 significant features is 95.7%. However, the accuracy of malicious websites is 88.2%. We should enhance our URL identification accuracy by adjusting the classifier or find more relevant features to improve our classification.

Reference

[1] C. Cortes and V. Vapnik, "Support-vector networks". Machine Learning No. 20, pp.273-297 (1995).

[2] C. M. Chen, J. J. Huang, Y. H. Ou, "Efficient suspicious URL filtering based on reputation". Journal of Information Security and Applications Vol. 20, pp.26-36 (2015)

[3] D. K. McGrath and M. Gupta, "Behind Phishing: An Examination of Phisher Modi Operandi". In Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats (LEET), San Francisco, CA, (2008).

[4] DMOZ - The Directory of the Web. https://www.dmoz.org/ (online)

[5] H. Choi, B. B. Zhu, and H. Lee, "Detecting Malicious Web Links and Identifying Their Attack Types". WebApps'11 Proceedings of

the 2nd USENIX conference on Web application development, pp.11-11, Berkeley, USA (2011)

[6] H. B. Kazemian and S. Ahmed, "Comparisons of machine learning techniques for detecting malicious webpages". Expert Systems with Applications, pp. 1166-1177 (2015).

[7] H. Xu, C. Caramanis, and S. Mannor, "Robust Regression and Lasso". IEEE Transactions on Information Theory, pp. 3561 - 3574 (2010).

[8] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs". In KDD '09 Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, USA, pp. 1245-1254 (2009).

[9] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song, "Design and Evaluation of a Real-Time URL Spam Filtering Service". 2011 IEEE Symposium on Security and Privacy, California, USA, pp. 447-462 (2011).

[10] L. Breiman, "Better Subset Regression Using the Nonnegative Garrote". Technometrics No.37, Pp.373-384 (1995).

[11] M. Aldwairi and R. Alsalman, "MALURLS: A Lightweight Malicious Website Classification Based on URL Features". Journal of Emerging Technologies in Web Intelligence (JETWI) Vol. 4, pp.128-133 (2012)

[12] M. Darling, G.Heileman, G. Gressel, A. Ashok, and P. Poornachandran, "A Lexical Approach for Classifying Malicious URLs". High Performance Computing & Simulation (HPCS), Alexandria,USA, pp.195-202 (2015).

[13] M. Egele, E. Kirda, and C. Kruegel, "Mitigating driveby download attacks: Challenges and open problems". In Proceedings of Open Research Problems in Network Security Workshop (iNetSec 2009), Zurich, Switzerland (2009).

[14] Malware - Clean MX. http://support.clean-mx.de/clean-mx/viruses.php (online)

[15] N. Provos, D. McNamee, P. Mavrommatis, K. Wang, and N. Modadug, "The ghost in the browser analysis of web-based malware". In: Proceedings of the first workshop on hot topics in understanding botnets, Cambridge (2007).

[16] R. Tibshirani, “Regression Shrinkage and Selection via the lasso”. Journal of the Royal Statistical Society Vol.58, pp.267-288 (1996).

[17] S. Salzberg, "On comparing classifiers: pitfalls to avoid and a recommended approach". Data Mining and Knowledge Discovery, Boston, USA, pp. 317-328 (1997).

[18] Security: https://sucuri.net/website-security/google-blacklisted-my-website (online)

[19] 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, China, pp.352-361 (2011).

[20] Zhou, Q., Song, S., Huang, G. and Wu, C., Efficient Lasso training from a geometrical perspective, Neurocomputing, Vol. 168, pp. 234-239 (2015).