Proceedings of the 2013 International Conference on Machine Learning and Cybernetics, Tianjin, 14-17 July, 2013

MALICIOUS WEBSITE DETECTION UNDER THE EXPLORATORY ATTACK

MANLIN WANG, FEI ZHANG, PATRICK P. K. CHAN

School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China
E-MAIL: 330567844@qq.com, 604660937@qq.com

Abstract:

Malicious websites provide a platform supporting diverse Internet crimes. They do not only steal the sensitive information but also let the hacker to control the computer of users. Malicious website detection with the machine learning technique achieves satisfying result. However, the characteristics of the malicious website may be modified to evade the detection. In this paper, the exploratory attack which misleads the decision of the classifier on the malicious samples by change the feature values with the minimum cost is discussed for malicious website detection. The costs of modifying features of malicious website domain are discussed. The attack model is used to attack the detection system using Support Vector Machine and Fisher Discriminant Classifier. The experimental results show that SVM is more robust than fisher discriminant classifier. Moreover, the vulnerable features are also discussed for the malicious website detection.

Keywords:

Malicious website detection; Adversarial Learning; Exploratory attack

1. Introduction

Malicious websites are designed for illegal purposes, for example, spam-advertising, malware propagating, and financial fraud through phishing. The malicious websites not only steal or damage the information from users, but also let the hackers to control the computers. It becomes a platform supporting diverse Internet crimes [1]. As a result, detecting the malicious sites to avoid the damage is an important problem.

One of the early solutions of malicious detection is the blacklist. A blacklist is a precompiled list containing IP address, domain name or URLs of known malicious websites [1]. The blacklist query service is a simple rule-based system which check whether a website to be visit is included in the blacklist. A website excluded in the blacklist is assumed to be safe. However, the response time to a new malicious website of the blacklist is slow. It requires a long period of time from discovering, verifying and updating a new malicious website to the blacklist. In the meanwhile, users are threatened by the new malicious websites. Moreover, the cost and the time required by creating a malicious website are low. Any new malicious website cannot be detected by blacklist-based system since they have different IP addresses and domain names. This drawback significantly downgrades the reliability of the blacklist-based protection system.

An adaptive and predictive protection based on machine learning methods were discussed [2,3,4,5,6]. Different from the blacklist which records all the discovered malicious websites, the machine learning methods classifies an unseen website according to the knowledge learnt from the dataset which contains the benign and malicious websites. A website is recognized as a malicious website if its properties are similar to the malicious websites. Therefore, the successful detection rate of machine learning methods is higher than the blacklist-based systems. Moreover, the machine learning methods do not require the manually update and is self-adaptive to the new malicious website.

To evade the detection, the malicious website creators try to modify the properties of the malicious website intentionally. This problem is called the adversarial learning [8]. The attacker and the defender adjust their strategies to beat the other side to increase its own performance. For example, if a website will be classified as a malicious one with high probability if it contains a keyword "banking" in the URL. The adversary will avoid containing this keyword in the URL. It will reduce the chance of classifying as malicious website.

To the best of our knowledge, no adversary attack has so far been discussed in malicious website detection. In this paper, we will firstly discuss the cost the modifying each feature in malicious website detection. Base on this information, an exploratory attack [9] which modifies the malicious website sample in the testing set to evade the detection is discussed. We assume the linear classifier is used in the detection system and the adversary has all information of the classifier. Modifying different features cost differently. A sample of malicious website will be modified to evade the detection with the lowest feature modification cost. This exploratory attack is implemented and evaluated using the Support Vector Machine (SVM) and Fisher Linear Discriminant (Fisher) classifiers. The robustness of the

978-1-4799-0260-6/13/$31.00 ©2013 IEEE

classifiers will be discussed.

The rest of the paper is arranged as follows. Section 2 provides the background of the malicious website detection and adversary learning. The exploratory attack model for malicious website detection is described in section 3 and section 4 discusses the experimental result. Finally, the conclusion will be given in section 5

2. Related Work

2.1 Malicious Website Detection

As the generalization ability of machine learning method is superior to the blacklist, the machine learning, i.e. the classification method, has been applied in malicious website detection. The features classifying malicious website can be categorized into two types. For the first type, the features can be retrieved without visiting the websites. Kan and Thi [2] proposed a method to analyze the string of URL of a website. A bag-of-words is applied to represent the tokens in the URL as the lexical features. They applied maximum entropy based learning on their features and achieved a 92% accuracy on the detection. McGrath and Gupta [12] analyzed the useful features to identifying the phishing websites. Those features include the IP address, the WHOIS records, the geographic information and the lexical features of the URL which are length, character distribution, and presence of pre-defined brand names. In [3], rather the information contained in URL or WHOIS, the information provided by external system, which are the Google's Page Rank and Google's Web page quality guidelines, are also applied into the system. With using the 18 features applied to the logistic regression classifier, the malicious website detection method is 97.3% accurate over a set of 2,500 URLs.

On the other hands, another type of features requires to visit the website. For example, the content of the website. The content-based features from the webpage, including whether iframes are "out of place", the presence of obfuscated javascript, and whether iframes point to known malicious sites, are extracted in [5]. Less expensive features are extracted to give a score of the webpage. Suspicious websites are further examined using the virtual machines. This work focuses on malware (drive-by-download) websites that are extremely harmful to computers. Phishing website, which is also one type of the malicious website, is classified using eight features: age of the domain, known images or logos, number of dots in the URL, whether the page is at suspicious, whether the page contain suspicious links, whether certain characters are present, whether the URL contains an IP address, and especially, their own calculated feature, TF-IDF, using a linear classifier [4]. The best performance of their classifier is 97% with 6% false positive or 89% with 1% false positive over a 100 phishing and 100 benign dataset. Visiting unknown websites may be potential harmful. However, analyzing the content of the website is more useful than only looking at the URL or WHOIS information usually.

Besides the single classifier, the Multiple Classifier System (MCS) is also applied to the malicious website detection [13]. The proposed method applies base classifiers in different feature subsets spaces, including foreign contents, script contents, DNS and URL information and exploit contents from different sources.

2.2 Adversarial Learning

Machine learning methods become a common technique in security applications, for example, spam mail detection [7]. An adversary who intentionally attempts to downgrade the performance of the system may exist in the security applications. This problem is called the adversarial learning. In the adversarial environment, the learning becomes complicated as the distributions of the training data and unseen samples are difficult due to the attack from the adversary [10]. For instance, the distribution of words appeared in email changes rapidly by adversary. The keywords which are highly predictive in the data set may not be useful anymore in next period of time.

The adversarial attacks can be categorized into causative attacks and exploratory attacks [8]. Causative attacks alter the training data in order to mislead the learning of a classifier. The benign and the malicious samples will be identified wrongly if the classifier did not learn the correct knowledge from the data. When the flipped labels reach a certain percentage, SVM classifier may reach a 50% error rate. [14] Alternatively, the training set is clean in the exploratory attacks. The adversary changes the testing samples to mislead the decision of the classifier. A typical example of the exploratory attack is the good word attack [9], which modifies spam messages by inserting a number of the words which always appear in legitimate mails, targeting on the spam-filtering system.

One point of views on the adversary learning is the game theory [11]. The adversary as the attacker and the classifier as the defender are two parties against each other in a game. Assume there are two classes of samples: + and – indicate the malicious and legitimate classes. Let $U(a; b)$ is the utility function, where a is the true class label of a sample and b is the output of the classifier. In the aspect of the classifier, the utility of the classifier is positive when the classifier correctly predicting the class labels of instances, which is $U_c(+; +) > 0$

or U_c (–; –) > 0. Otherwise, the utility is negative (U_c (+;–) < 0 or U_c (–; +) < 0). On the other sides, as the goal of adversary is to mislead the decision of the classifier, the utility is positive when U_c (+;–) < 0 , U_c (–; +) < 0. Otherwise, it is negative. Both the adversary and the classifier try to maximize their utility functions.

3. Exploratory Attack for Malicious Website Detection

We apply the attack model in [11] as attack framework in this paper. The utility of an successful attack for the adversary is

$$Ua = Ga(+;-) - \sum_{i=1}^{n} C(i) \tag{1}$$

where n is the number of features, G_a (+;–) represents the gain for the adversary when the classifier identifies a malicious sample (which is + class) as benign (– class), and *C(i)* is the cost of modifying the *i*th feature to the desired value. *C*(*i*) is defined as

$$C(i) = \Delta x_i * Cost(i) \tag{2}$$

where Δx_i is the difference between the desired feature value and its original value of the *i*th feature, and *Cost*(*i*) is the cost of modifying each unit value of the feature the *i*th feature.

The costs of the 15 features in malicious website detection are discussed in this paper. The 15 features categorized into three types (lexical features of URL, contend related features and host-based features) and their costs are listed in Table 1.

TABLE1. COST OF FEATURE MODIFICATION

i	**Feature (f_i)**	**Cost(*i*)**
1	Number of tokens in domain	1
2	Average length of domain tokens	1
3	Existence of @ in URL	4
4	Existence of special words	1
5	Number of % in URL	1
6	Length of title	1
7	Relevancy with title	4
8	Webpage size	3
9	Compressed webpage size	3
10	Compression rate	3
11	Existence of robot	2
12	Type of server	1
13	Loading time	2
14	Global rank	4
15	Local rank	4

The first five features are lexical features of URL. Feature 1 and 2 are related to the tokens of URL, which are the strings in URL composed by numbers (0-9) and letters (a-z and A-Z) separated by other symbols. Since the cost for modifying URL is relatively low, the cost level for these two features is 1. The words between http:// and @ will be ignored by the browser but not by the user. For example, a malicious website may use an URL like *http://www.google.com.hk@www.malicious.com* to cheat users to visit it. As a result, the cost of removing @ from the URL (f_3) is very high. For feature 4, the special words refer to: "confirm", "banking", "signin", "ebayisapi", "login", "webscr", "account", "secure", "secur", "notif", "log", "click", "verify", "update", "inconvenien" and "pay". These words frequently appear in the URL of malicious website especially phishing. However, an easy way which adding some symbols before or after the special word (e.g. change "account" to "+account") can change the feature value but users still can read the same special word. Some browsers interpret "%+X" where X is a integer as a letter. Number of % in URL indicates the potential letters hidden in the URL. Similar to other lexical features, the cost of change this features is not high.

On the contrary, modifying the content of malicious websites (f_i where i = 6 – 11) is generally more expensive. The title length is defined as the number of characters in the title of the webpage. As changing a title is costless, the cost of the feature 6 is low. The feature 7 "relevancy with title" means how much the webpage content relates to the keywords in its title. Changing the content of a website may cost a lot for the adversary. The feature webpage size (f_8) denotes the number of bytes of the whole webpage. If a spam-advertising website needs to reduce the value of this feature, it has to use less or lower quality photographs to promote the products, which is less attractive to users. Similar the cost of changing compressed webpage size and rate (f_9 and f_{10}) is as the same as feature 8. Web robot is a program which automatically traverses a website retrieving information. The robots appear frequently in malicious websites. The cost of changing the feature 11 (Existence of robot) is not high as the functionality of robots can be replaced by other technologies.

We also evaluates some host-based features (f_i where i = 11 – 15). The cost of changing the feature 12 "type of server" (e.g. Apache or Window NT) is relatively low for adversaries. Loading time (f_{13}) is defined as the number of seconds for a user to load the webpage. Changing this feature implies change the size of website and the bandwidth of the server. Therefore, the cost is not low. The two ranks are given by alexa [19]. Global rank (f_{14}) represents the rank of the website in the world, while local rank (f_{15}) for the rank of the website in its country. The cost for changing the ranks is very high since it requires beating other websites in the alexa ranking in term of traffic of website visiting.

The objective of adversary is to maximize the utility function (1) for each malicious sample in the testing set. We assume the term $G_a(+;-)$ is the same for every samples, which means the reward of successful misleading the classifier on each malicious sample is the same. Thus, the malicious samples which already classified wrongly by the classifier will not be considered in the attack. The problem can be formulated as an optimization problem for each malicious sample classified correctly shown as follows:

$$\min_{\Delta x} \sum_{i=1}^{k} \left(\Delta x_i \times COST(i) \right), \ s.t. \ G_a(+;-) \qquad (3)$$

where Δx is $\{\Delta x_1, \Delta x_2, \ldots \Delta x_k\}$ and k is the number of features. For each malicious sample classified correctly by the classifier, we would like to change the feature value with minimum costs to change the decision of the classifier on the malicious sample.

4. Experiment

In this part, we choose SVM and linear discriminant classifier (FISHER) as our target and compare their robustness by calculating the cost for attacking them. And we will discover some features that are more preferable by the attacker.

4.1 Malicious Website Dataset Preparation

The malicious websites contain three different types in our data set: phishing, malware and spamming websites. Phishing websites are those pretend to be e-bank or famous e-commerce websites in order to cheat users to submit their information such as account and password. Malware indicates those websites that automatically download malicious software into users' computers and thus take the control of the computers. Spamming websites pop spam messages such as advertisement to users. The phishing, malware and spamming websites are collected from the blacklist providers [15, 16,17] randomly. The benign URLs are obtained randomly from *randomwebsite.com* [18].

The 15 features mentioned in Table 1 are collected for each URL. The 15 features can be categorized into three types: lexical features, the feature related to the malicious website content and host-based features. Type of server, compressed webpage size and compression rate are collected from [20], while global rank and local rank are obtained from [19]. [21] provides the information of length of title, relevancy with title, webpage size, existence of robot and loading time. The rest of five lexical features, which are number of tokens in domain, average length of domain tokens, existence of @ in URL, existence of special words and number of % in URL, are extracted from analyzing the string of URL.

The samples with missing features are removed. Finally, the dataset contains 2199 phishing, 525 malware and 1530 spamming websites. In this paper, we only focus on 2-class problem, which means a sample is either classified as benign or malicious website. As a result, all kinds of malicious websites are combined as one class. The dataset used in this paper has 3362 benign samples and 4254 malicious samples.

4.2 Experimental Result of Exploratory Attack for Malicious Website Detection

The experiment is executed ten times individually. The inputs of all samples are normalized to [0, 1] to eliminate the effect of different ranges of values. The formula $(f_{ij} - \min_i) / (\max_i - \min_i)$ is used to normalize a value (f_{ij}) of the i^{th} feature in the j^{th} sample, where $\min_i$ $(\max_i)$ is the minimum (maximum) value of the i^{th} feature. 50% of the dataset is selected randomly as a training set and the rest of samples are grouped as the testing set for each experiment. Support Vector Machine (SVM) with the linear kernel and Fisher Linear Discriminant (Fisher) classifiers are used in the Malicious Website Detection. In each experiment, SVM and Fisher are trained by using the training samples. The trained classifiers are attacked by the model mentioned in Section 3. The testing set contains 2127 malicious samples. In average, SVM and Fisher classify 1811.9 and 1714.7 out of 2127 malicious samples correctly. The attack method focuses on these correctly classified samples. The results of the attack are discussed and analyzed in this section.

Figure 1 shows the average successful attack rate with different maximum costs on the ten independent runs. Y-axis indicates the number of successful attacked samples and x-axis is the maximum attack costs used by the attack model for each sample. The thick and thin lines represent SVM and Fisher. For example, if the adversary allows 0.2 maximum cost to attack each sample, around 50% of malicious samples can mislead SVM. Figure 1 shows that using the same maximum cost, more samples can be attacked if the detection system using Fisher comparing with SVM. For example, when the maximum cost is 0.1, only less than 10% of samples can mislead SVM but Fisher classifies more than 60% of malicious samples. Moreover, only using a small value of cost, i.e. 0.05, around 25% of samples are successfully attacked for Fisher. However, only around 5% of samples can be attacked for SVM. The same experimental result is shown in the figure 2 which is the density of the successful attack samples against different maximum attack cost. It shows that using 0.35 maximum cost can mislead the

decision of Fisher on all malicious samples. On contrast, SVM classifies all the samples only if the maximum cost is around 0.7. The results indicate that the performance of SVM is more robust than Fisher under the exploratory attack mentioned in section 3, which means the adversary needs to pay more cost to attack the same amount of samples when SVM is used in the malicious website detection system comparing with Fisher. The decision plane of SVM maximizes the distance in the margin which is a corridor between two classes. The adversary needs a large Δx to attack a malicious sample since the distance of between the samples and the decision plane of SVM is large. In general, from formula (3), an attack requires a larger Δx causes a larger cost. It may explain why SVM is more robust than Fisher under the proposed exploratory attack.

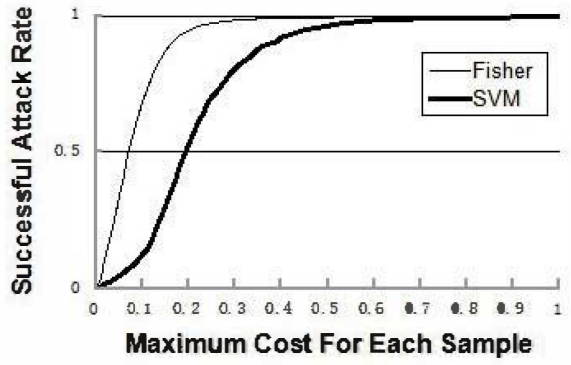

Figure 1. Successful Attack Rate with Different Maximum Cost per Each Sample

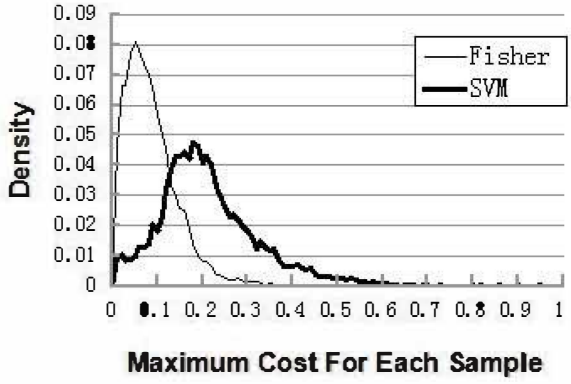

Figure 2. Distribution of Successful Attack Samples with Different Maximum Cost per Each Sample

Table 2 shows the modification information of features of attack sample for both SVM and Fisher when all malicious samples are attacked. A row in the table represents a feature. Each classifier has two columns. The first column (%) indicates the percentage of all attack malicious samples which modify the feature and another column (Rank) shows the rank of the percentage among all features. The rank with smaller value means it has a large percentage. The last column indicates the average rank of Fisher and SVM.

Although the cost of changing website size and compressed website size (f_8 and f_9) is not low (cost is 3), both features are changed frequently in the attack. The attack changes the value of these two feature in more than 75% samples. It may because that the website sizes of benign and malicious website are different. Changing this feature slightly can mislead the classifier easily. Another vulnerable feature is the feature 1 "the number of tokens in domain". The malicious websites usually contain more tokens in the URL than benign ones. As the attack cost of feature 1 is only 1, it also attracts the attack from the adversary. The rank of the type of server is the largest among the 15 features because benign and malicious websites may use the same server. Similarly, the existence of robot does not indicate if a website is malicious or not. Attacking this feature is not efficient.

TABLE 2. FREQUENCY OF ATTACK

i	Feature (f_i)	Fisher		SVM		Avg.
		%	Rank	%	Rank	Rank
1	Number of tokens in domain	73.9	3	85.2	1	2.0
2	Average length of domain tokens	67.3	6	68.7	6	6.0
3	Existence of @ in URL	58.3	11	47.0	15	13.0
4	Existence of special words	57.8	12	54.8	12	12.0
5	Number of % in URL	64.1	9	57.2	10	9.5
6	Length of title	67.7	5	69.4	5	5.0
7	Relevancy with title	56.5	13	54.9	11	12.0
8	Webpage size	76.7	2	78.1	4	3.0
9	Compressed webpage size	80.0	1	83.0	2	1.5
10	Compression rate	65.8	7	65.6	7	7.0
11	Existence of robot	55.7	14	54.2	13	13.5
12	Type of server	54.1	15	52.6	14	14.5
13	Loading time	72.0	4	82.7	3	3.5
14	Global rank	63.9	10	62.6	9	9.5
15	Local rank	65.0	8	64.0	8	8.0

In summary, according to our experiment, SVM classifier is generally more robust to against the exploratory attacks discussed in this paper than the fisher discriminant classifier. Moreover, the features related to the website sizes, loading time and the number of tokens are more vulnerable than other features. To reduce the damage of the attack, the smaller weight should be assigned to these features when constructing the classifier.

5. Conclusions

The properties of the malicious website may be modified to evade the detection. In this paper, we discuss and perform the exploratory attack on malicious detection systems. The exploratory attack misleads the decision of the classifier on the malicious samples by change the feature values with the minimum cost. We firstly discuss the cost of modifying features of malicious websites. In general, the cost of changing the website content related features are highest. The cost of lexical features is lower than the host-based features. The attack model with this cost information is applied to attack the malicious detection system with SVM and fisher discriminant classifier. The experimental results show that SVM is more robust than fisher discriminant classifier in term of less samples can be attacked using the same attack cost. Moreover, the vulnerable features are also discussed. The features related to the website sizes, loading time and the number of tokens are the most vulnerable among the features used in the experiment.

Acknowledgments

This work is supported by National Natural Science Foundation of China (61272201, 61003171 and 61003172), a Program for New Century Excellent Talents in University (NCET-11-0162) of China and "the Fundamental Research Funds for the Central Universities" No.201325 of South China University of Technology.

References

[1] Justin Tung Ma , Lawrence K. Saul , Stefan Savage , Geoffrey M. Voelker: Learning to Detect Malicious URLs. ACM Transactions on Intelligent Systems and Technology (TIST), Volume 2 Issue 3, 2010.

[2] Min-Yen Kan and Hoang Oanh Nguyen Thi: Fast Webpage Classification Using URL Features. In Proceedings of the 14th ACM international conference on Information and knowledge management (2005), pp. 325-326, 2005.

[3] Sujata Garera, Niels Provos, Monica Chew, and Aviel D. Rubin. A: Framework for Detection and Measurement of Phishing Attacks. In Proceedings of the 2007 ACM workshop on Recurring malcode (2007), pp. 1-8, 2007.

[4] Yue Zhang, Jason Hong, and Lorrie Cranor. CANTINA: A Content-Based Approach to Detecting Phishing Web Sites. In Proceedings of the 16th international conference on World Wide Web, Pages 639-648, 2007.

[5] Niels Provos, Panayiotis Mavrommatis, Moheeb Abu Rajab, and Fabian Monrose: All Your iFRAMEs Point to Us. In Proceedings of the 17th conference on Security symposium (2008), pp. 1-15, 2008.

[6] D. J. Guan, Chia-Mei Chen, and Jia-Bin Lin: Anomaly Based Malicious URL Detection in Instant Messaging, Workshop on Information security, JWIS, 2009.

[7] Han Xiao, Huang Xiao: Adversarial and Secure Machine Learning, http://www.sec.in.tum.de/adversarial-and-secure-machine-learning

[8] Marco Barreno, Blaine Nelson, Russell Sears, Anthony D. Joseph, J. D. Tygar: Can Machine Learning Be Secure? In Proceedings of the ACM Symposium on Information, Computer, and Communication Security, Pages 16-25, 2006.

[9] Daniel Lowd, Christopher Meek: Good Word Attacks on Statistical Spam Filters. In Proceedings of the Second Conference on Email and Anti-Spam (CEAS), 2005.

[10] Amir Globerson, Sam Roweis: Nightmare at Test Time: Robust Learning by Feature Deletion. In Proceedings of the 23rd international conference on Machine learning, Pages 353-360, 2006.

[11] Nilesh Dalvi, Pedro Domingos, Mausam, Sumit Sanghai, Deepak Verma: Adversarial Classification. In Proceedings of the 10th International Conference on Knowledge Discovery and Data Mining, Pages 99-108, 2004.

[12] D. Kevin McGrath and Minaxi Gupta: Behind Phishing: An Examination of Phisher Modi Operandi. In Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats, Article No. 4, 2008.

[13] Van Lam Le, Ian Welch, Xiaoying Gao, Peter Komisarczuk: A Novel Scoring Model to Detect Potential Malicious Web Pages. 2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications, page 254-263, 2012.

[14] Han Xiao, Huang Xiao, Claudia Eckert : Adversarial Label Flips Attack on Support Vector Machines. In Frontiers in Artificial Intelligence and Applications, page 870-875, 2012.

[15] http://www.spamcop.net/spamstats.shtml

[16] http://www.malwaredomainlist.com/mdl.php

[17] http://www.phishtank.com/

[18] http://www.randomwebsite.com/

[19] http://www.alexa.com/

[20] http://www.gidnetwork.com/tools/gzip-test.php

[21] http://www.metachecker.net/