Two-Phase Malicious Web Page Detection Scheme Using Misuse and Anomaly Detection

Suyeon Yoo and Sehun Kim
Graduate School of Information Security, Korea Advanced Institute of Science and Technology (KAIST), 291 Daehak-ro, Yuseong-Gu, Daejeon 305-701, Republic of Korea
yoosuyeon@kaist.ac.kr, shkim@kaist.ac.kr

Abstract

Misuse detection method and anomaly detection method are widely used for the detection of malicious web pages. Both are based on machine learning. Misuse detection can detect known malicious web pages, but it cannot detect new ones. In contrast, anomaly detection can detect unknown malicious web pages, but it has a high false positive rate. In order to achieve a high detection rate through precisely detecting known and unknown malicious web pages, we propose a two-phase detection scheme. In the first phase, the misuse detection model is built based on the C4.5 decision tree algorithm, which allows known malicious web pages to be detected. In the second phase, the anomaly detection model with a one-class support vector machine is used to detect new types of malicious web pages. The experimental results show that our proposed method has significantly higher malicious web page detection rate than conventional ones with the expense of slightly higher false positive rate.

Keywords: Anomaly detection, machine learning, malicious webpage, misuse detection

1. Introduction

With the development of computer networks, more services have become available on the Internet and many people can access Internet services via web applications. A variety of web applications provide convenient services to users, such as searching for information, using online commerce, and communicating through social network services. However, as these activities that support people in their daily lives become intertwined with the Internet, vulnerable web pages and applications are becoming increasingly attractive targets for malicious attacks and data theft.

Web pages in web applications are commonly described using computer languages such as HTML code and JavaScript [10]. While these computer languages support web application developers, the source codes written in these languages can be easily modified by attackers. Simple modification of source codes can create new types of malicious web pages and can have numerous victims. That is, the users visiting the vulnerable web pages can become victims of attacks. The attacker can obtain the user’s critical information from the computer and use the infected computers as a path to compromise other computers connected to the same network. Therefore, precise detection of malicious web pages and blocking new types of malicious pages are very important.

Most techniques that have been used to defend users against malicious web pages can be classified into three categories: blacklisting, dynamic analysis, and static analysis. Among these three approaches, the most common protection approach is blacklisting: previously identified lists of known malicious properties such as URLs, domain names, and IP addresses can be blocked from users’ access [1], [3], [4], [5]. Although blacklisting is a simple and
precise detection method for identified malicious web pages, it cannot evaluate new web pages that have not yet been blacklisted, even though they are malicious [3].

The dynamic approaches attempt to capture unusual behaviors, such as launching attacks and destroying user computers, when the page is explored in a controlled environment, such as client honeypots and virtual machines [3], [6], [7], [9], [11], [12]. Because dynamic analysis techniques scrutinize the web-based scripts associated with web pages to detect if the page is malicious, these techniques have a high detection rate but are resource intensive. In addition, analyzing web pages in controlled environments requires a long time.

In order to overcome the limitations of blacklisting and dynamic analysis, there have been studies based on static detection techniques that use machine learning to detect malicious web pages [2], [5], [8], [10], [18], [19]. By learning previously observed patterns of normal and malicious web pages, the machine can extract the static features and make predictions during real tests. In machine learning-based techniques, the detection time is faster than in dynamic analysis techniques, but the detection accuracy is lower than in dynamic analysis [3], [8]. Although previously proposed static studies primarily exhibited low detection rates, the studies of detection methods using machine learning are increasing because they are appropriate for large-scale data analyses. Therefore, this study proposes a malicious web page detection method based on machine learning.

In general, machine learning-based detection approaches fall into two major categories: misuse detection and anomaly detection. The misuse detection approaches, commonly known as signature-based detection, use a training dataset that includes patterns of normal web pages and patterns known to occur in malicious web pages. The anomaly detection methods train machines with normal patterns of web pages and look for deviations from these patterns.

For misuse detection, Ma et al. [19] proposed a detection method that uses only the lexical and host-based features of URLs: their study focused on lightweight URL classification. Canali et al. [2] and Choi et al. [8] proposed fast filtering systems based on misuse detection, and these filtering systems are allocated before dynamic detection systems such as client honeypots. While both studies [2] and [8] reused a number of features from [19] and [10], respectively, they introduced effective detection frames using misuse detection-based classifiers. In order to locate more suitable features for classifying web pages, Hou et al. [10] focused on web content features and the two studies by Eshete et al. [5], [18] extracted features that included the static aspects, dynamic aspects, and metadata of URLs. These misuse detection approaches work reliably on well-known malicious web pages, but have a clear disadvantage because they are not capable of detecting unknown types of malicious web pages [17].

There have also been a few studies in malicious web page detection that address the anomaly detection methods [23], [24]. Ying et al. [23] proposed a web phishing page detection using an anomaly approach. By identifying new types of web pages, the detection performance was good but it was only appropriate for web phishing attacks. Marco et al. [24] introduced a web attack detection method that combined anomaly detection with emulation to identify malicious JavaScript code. Although anomaly detection can detect new malicious web pages, it cannot identify normal web pages that are unusual [17].

In order to resolve the disadvantages of the misuse detection and anomaly detection methods, a novel framework is proposed that hierarchically integrates a misuse detection model and an anomaly detection model. This study is a pioneering attempt to combine misuse and anomaly detection methods for the detection of evolving malicious web pages, even though similar approaches exist in the traditional research field of network intrusion detection. In this research, the new hybrid detection method has a two-step hierarchy. The first step is
misuse detection and it uses a supervised learning algorithm. The second step is anomaly
detection and it uses an unsupervised learning algorithm. Through organizing the hierarchical
scheme, the well-known malicious web pages can be detected using the misuse detection
model and the unknown malicious web pages, which are obtained from the suspicious web
pages that were identified in the first detection component, can be detected using the anomaly
detection model.

In addition, the proposed detection method uses an unsupervised support vector machine
(SVM), so called one-class SVM, as the classifier in the anomaly detection model. This one-
class SVM has garnered significant interest and has become one of the most developed
methods in anomaly detection. In the previous research, SVMs have been used as supervised
learning classifiers in the misuse detection model that needs labeled training data sets
including patterns of known malicious web pages and normal web pages. However, other
anomaly detection models that use the one-class SVM to detect malicious web pages have not
been reported.

The proposed method was evaluated through experiments using 171 features extracted web
content and two classifiers (C4.5 decision tree (DT) for the misuse detection model and one-
class SVM for the anomaly detection model). The experiment results demonstrate that the
detection rate of the proposed method is 98.9%. This rate is better than that of the
conventional misuse detection model using a DT and the anomaly detection model using the
one-class SVM.

This paper is organized as follows. Section 2 explains the features of the proposed method.
Section 3 describes the proposed algorithm in detail. The algorithm is evaluated in Section 4.
Finally, the study is summarized and the paper is concluded in Section 5.

2. Features

The proposed detection algorithm uses a machine learning technique and it requires the
features that can distinguish malicious web pages and benign web pages. The proposed
method uses the features presented in the previous study [10] to detect malicious web pages
using machine learning. Because only dynamic HTML webpage features are used to classify
the web pages, the features used in the earlier research [10] can satisfy these criteria. All
features used in this paper are categorized according to the usage of dynamic HTML
knowledge [10]. There are three feature groups: HTML document features, JavaScript
features, and ActiveX features.

A. HTML Document Features

Document features can be easily extracted from HTML documents; furthermore, document
features have powerful discriminating abilities [10]. Therefore, they are used as features in
machine learning-based detection methods [8], [10]. In the HTML code document, the
number of lines, size of IFRAMEs, number of words per line, and average length of words
can be features that present the essential characteristics of malicious web pages.

B. JavaScript Features

Because the number of attacks using JavaScript functions is increasing, the JavaScript
features must be considered. JavaScript features are numerical features: the number of uses of
each JavaScript function. These functions include escape(), unescape(), eval(), ubound(), and
exec(); these are commonly used to inject malicious codes into web pages. In addition, there
are hundreds of JavaScript functions that can be used as features [10]. According to the use of
the JavaScript features, malicious web pages that use JavaScript functions are detected by
verifying how often these malicious functions are used [8]. In this study, 154 JavaScript functions used in the earlier study were selected as the level of significance [10].

C. ActiveX Features

ActiveX objects are functions that used on web pages because they can enhance the web browsing experience. However, attackers can embed their malicious code by inserting it into the ActiveX controls, which is a small program created by the ActiveX objects; for example, Adodb.Stream is used by attackers to download files from the Internet [10]. Because web pages can be infected by malicious codes through the vulnerability of the ActiveX features, the ActiveX objects are considered when detecting malicious web pages. The ActiveX features used in the aforementioned study are also used in this study [10]. The number of selected features used in all experiments is listed in Table 1.

<table>
<thead>
<tr>
<th>Table 1. Selected features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature name</td>
</tr>
</tbody>
</table>
| HTML document features | 9 | 1. Number of words  
2. Number of words per line  
3. Line count  
4. Average word length  
5. Null space count  
6. Number of delimiters  
7. Distinct word count  
8. Whether the ‘script’ tag is symmetric  
9. The IFRAMEs size |
| JavaScript features | 154 | Number of executions for each Java function |
| ActiveX features | 8 | Number of executions for each ActiveX object |

3. Proposed Two-Phase Detection Method

A malicious web page detection method that includes the misuse detection and anomaly detection approaches is proposed. In the introduction, two methods to hierarchically combining the misuse detection module and anomaly detection module were explained briefly: the framework of the anomaly detection module followed by the misuse detection component and the framework of the misuse detection component followed by the anomaly detection component. The first method locates suspicious web pages using anomaly detection, and then it applies the misuse detection method to detect malicious web pages from the suspicious web pages identified in the first step. In this approach, the anomaly detection module should correctly detect most malicious web pages, because overlooked malicious web pages cannot be detected by the following misuse detection module [13]. However, if the anomaly detection algorithm is implemented before the misuse detection algorithm, the anomaly detection module could have a low detection rate through overlooking malicious web pages that have well known signatures but are similar to benign web pages because the anomaly detection algorithm is based on the hypothesis that malicious web pages have different characteristics to normal web pages. Therefore, it is difficult to obtain a high detection rate for known malicious web pages in the anomaly detection module.

In order to increase the detection rate, the proposed malicious web page detection method implements the misuse detection module followed by the anomaly detection module. This approach has been used for intrusion detection [13], [14], [15], [27]; however, there is no published research for malicious web page detection using this hybrid detection approach. The proposed framework of the hybrid detection is illustrated is Fig. 1. As shown in Figure 1, every web page is analyzed first by misuse detection module. The misuse detection module
applies the decision tree algorithm to detect the malicious web pages that have well-known characteristics by matching the patterns of malicious web pages. Then, the web pages that are not classified as known malicious web pages are provided to the anomaly detection module in order to detect new types of malicious web pages using the one-class SVM. The misuse detection module and anomaly detection module are explained in detail below.

**Step 1. Classification using the J48 Algorithm for the Misuse Detection Module**

In order to detect the known and unknown malicious web pages, the proposed detection method initially classifies the known malicious web pages using the misuse detection module, which is built using a decision tree algorithm. The decision tree [21] algorithm recursively partitions the input dataset into smaller subsets based on the features of the dataset. In order to build the misuse detection module, the J48 algorithm was used; it is known as a WEKA implementation of the C4.5 algorithm [20], [22]. This algorithm is one of the most well-known and widely used decision tree algorithms for machine learning.

Figure 1 presents a graph that illustrates the process of detecting known malicious web pages in the misuse detection module. As seen in the figure, the known malicious web pages can be detected, whereas the unknown malicious web pages are randomly classified in the misuse detection module. Without the classified known malicious web pages, the other items including normal web pages and unknown malicious web pages are considered uncertain items, and they are used as input in the anomaly detection module.

**Step 2. One-Class Support Vector Machine for Anomaly Detection Module**

In the second stage of the hybrid detection system, the uncertain items (including new malicious web pages and benign web pages) are classified as malicious web pages or benign web pages using the anomaly detection. The anomaly detection module uses the one-class support vector machine (1-class SVM) that was proposed by Scholkorft et al. [16]. This 1-class SVM is a modified classification algorithm from the general SVM and it detects outliers that have different attributes to the trained normal data. Therefore, new types of malicious web pages can be detected using the anomaly detection module because the 1-class SVM classifier is trained with the feature set of benign web pages. The process of detecting unknown web pages in uncertain items is also illustrated in Figure 1.
4. Evaluations

The proposed detection method that integrates the misuse detection module and anomaly detection module is evaluated. In this section, the experimental environment is described and the experiment results are compared with results from other detection methods.

A. Experimental Environment

There were 3,588 normal web pages and 345 malicious web pages collected in order to verify the performance of the proposed detection method. In order to collect these normal web pages and malicious web pages, RafaBot, which is a web crawler, and JSUnpack [26], which is a tool for analyzing malicious code, were used. All experiments were performed using WEKA (version 3.6.6, DEVELOPER, COUNTRY) [20].

According to the features described in Section 2, the number of selected features that were used in all experiments is 171. In order to test the performance, the proposed hybrid detection method is compared with the conventional decision tree and one-class SVM. The performance measurements are the detection rate and false positive rate. The detection rate is the rate of detected malicious web pages from the total number of malicious web pages; the false positive rate is the rate of web pages misclassified as malicious from the total number of normal web pages. The proposed detection method that uses the hybrid approach focuses on achieving high detection rates.

B. Experimental Results

The purpose of the proposed method is to detect malicious web pages while achieving a high detection rate. In order to increase the detection rate, a hybrid detection method that combines the misuse detection and anomaly detection components was introduced. The experiment results of the proposed method are presented in TABLE III. The detection rate of the malicious web pages was 98.9% and the false positive rate was 30.5%. According to the classification results, the proposed method can detect almost all malicious web pages.

C. Comparison with Other Detection Methods

In this section, the performance of the proposed detection method is compared with the other detection methods, including the single misuse detection method using the decision tree algorithm and the single anomaly detection method using a one-class SVM. TABLE IV provides an overview of the experiment results in comparison with the other detection methods.

The single misuse detection using the decision tree has a detection rate of 86% and a false positive rate of 0.8%; the single anomaly detection using a one-class SVM has a detection rate of 87.5% and a false positive rate of 27.8%. The lowest false positive rate was exhibited by the decision tree detection method and the proposed detection method had the highest detection rate among the three detection methods.

From the experiment results, the performance of the proposed hybrid detection method can be verified. Therefore, it is clear that the proposed approach performs well for detecting new and unseen malicious web pages, but its false positive rate is relatively high because the input web pages are classified in two phases. That is, the uncertain items in the anomaly detection component have a higher similarity than the
original input items. Although the false positive rate of the proposed detection method is relatively high, the proposed hybrid method has the highest detection rate.

5. Conclusion

As web-based services are increasingly provided, attackers use these web pages to inject malicious code and attack users. A detection method for these malicious web pages was proposed in this study. In particular, the proposed detection method can be used for known malicious web pages, as well as new types of malicious web pages. The proposed malicious web page detection method is a hybrid method based on machine learning techniques. This hybrid method is composed of a misuse detection module and an anomaly detection module in order to solve the disadvantages of both detection methods. The misuse detection approach can detect known attacks well and the anomaly detection approach can detect new types of malicious web pages well.

The experiment results of the proposed detection method exhibited a significantly improved detection rate of 98.9%. However, the proposed method had a relatively high false positive rate of 30.5% when compared with the single misuse detection and single anomaly detection methods. This result was caused by the increments of input data similarity in the anomaly detection component. Future work will focus on lowering the false positive rate.

References


Table 2. Experimental environment

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of normal web pages</td>
<td>3588</td>
<td></td>
</tr>
<tr>
<td>Number of malicious web pages</td>
<td>345</td>
<td></td>
</tr>
<tr>
<td>Data collection tool</td>
<td>RafaBot</td>
<td></td>
</tr>
<tr>
<td>Classification tool</td>
<td>WEKA 3.6.6</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Classification results

<table>
<thead>
<tr>
<th></th>
<th>Predicted normal</th>
<th>Predicted malicious</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real normal</td>
<td>0.695</td>
<td>0.305</td>
</tr>
<tr>
<td>Real malicious</td>
<td>0.011</td>
<td>0.989</td>
</tr>
</tbody>
</table>

Table 4. Comparison of the propose method results with the single misuse detection and single anomaly detection methods

<table>
<thead>
<tr>
<th></th>
<th>Proposed method</th>
<th>Decision tree</th>
<th>One-class SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detection rate</td>
<td>98.9%</td>
<td>86%</td>
<td>87.5%</td>
</tr>
<tr>
<td>False positive rate</td>
<td>30.5%</td>
<td>0.8%</td>
<td>27.8%</td>
</tr>
</tbody>
</table>


Authors

Suyeon Yoo

She received the B.S degree in industrial engineering from Pusan National University, Busan, Korea, in 2011 and the M.S degree in industrial and systems engineering from the Korea Advanced Institute of Science and Technology, Daejeon, Korea, in 2013. Currently, she is working toward the doctoral degree in information security within the Graduate School of Information Security. Her research interests are security issues such as malicious behavior detection in webpages and domain name server.

Sehun Kim

He received the B.S. degree in physics from Seoul National University, Seoul, Korea, in 1972, and the M.S. and Ph.D. degrees in operations research from Stanford University in 1978 and 1981, respectively. In 1982, he joined the faculty of KAIST. He has published a number of papers in IEEE Trans. on Vehicular Technology, Computer Networks, Telecommunication Systems, IEICE Transactions on Communications, International Journal of Satellite Communications, and Journal of KIISC (Korea Institute of Information Security and Cryptology). He served as the chief editor of the Journal of KIISC from 1990 to 1993.