Optik 127 (2016) 345–351

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.de/ijleo

A minimum enclosing ball-based support vector machine approach for detection of phishing websites

Yuancheng Li, Liqun Yang*, Jie Ding

School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, PR China

ARTICLE INFO

Article history:
Received 27 December 2014
Accepted 12 October 2015

Keywords:
Phishing website detection
DOM tree
Topological feature
Web crawler
Features extraction
BVM classifier

ABSTRACT

In this paper, a novel approach based on minimum enclosing ball support vector machine (BVM) to phishing Website detection is proposed, which aims at achieving high speed and high accuracy for detecting phishing Website. In order to enhance the integrity of the feature vectors, we first perform an analysis of the topology structure of website according to the DOM tree and use the Web crawler to extract 12 topological features of the website. Then, the feature vectors are detected by BVM classifier. Compared with the general SVM, this method has relatively high precision of detecting, and complements the disadvantage of slow speed of convergence on large-scale data. The experimental results show that the proposed method has better performance than SVM, and further validate the validity and correctness of our scheme.

1. Introduction

Phishing, just as its name implies is a way of internet fraud attacks, which send a mass of fraudulent E-mails and post forged web sites. After diddling the victims' trusts, the personal and private information, such as bank accounts, usernames, and passwords appear to be accessed through phishing scams. In modern society, as sales channels of various industries rely strongly on the Internet, phishing websites are being developed in a variety of forms and it is likely to disrupt every field it touches. An APWG (Anti-phishing Working Group) report shows that there were around 60,000 phishing websites appeared in email client or the websites were visited regularly by netizens [1]. Phishing scams have now overtaken the malicious URL injection to become the primary threat of web-surfing security.

In previous studies of phishing detection, there were many automatic phishing detection methods that could be used to identify phishing websites accurately. A new approach is proposed to detect phishing websites through prioritizing common words by search engines, which estimates domain name and content relevance to check if the web page is legal, and the recognition rates of this method can reach 97% [2]. Kirda et al. [3] proposed the use of browser plug-in to analyze whether the user submission information is sensitive information in real time, and then set prewarning to users. Meanwhile, many antiphishing tools in the form of browser plug-in emerged in the Internet, such as phishing websites filter in Internet Explorer, Firefox browser from Google, SpoofGuard developed by Stanford University [4–6], and these tools are based on blacklist, whitelist, and heuristic analysis. When a visited website is judged to a phishing website, the tools will pop-up prompt box to warn users. Weiwei et al. [7] proposed the use of clustering ensemble method to aggregate multiple cluster algorithms, which can develop an automatic categorization system to automatically classify phishing websites. Dhamija et al. [8] proposed dynamic security skins method that is based on server schemes. Topkara et al. proposed a novel scheme 'ViWiD', which is an integrity check mechanism based on visible watermarking of logo images and its implementation for mitigating phishing attacks [9]. Ying and Xuhua [10] analyzed the properties of web page based on structural DOM model, and utilized support vector machine (SVM) in detecting phishing page, but this method has some limitation when dealing with image. Consider the spread of phishing website Abu-Nimeh et al. [11] proposed a feature extraction method that mainly extract the feature of the emails with phishing webpages, then evaluated the effects of six machine learning algorithms in terms of classification. Although this method improves the detection precision in detecting phishing web sites, it uses a single web page to extract features, and therefore, it is easy to be deceived by phishing website designers.Due to new phishing websites emerge in an endless stream, and they have the characteristics of cost-effective

* Corresponding author at: School of Control and Computer Engineering, North China Electric Power University, 2 Beinong Road, Huilongguan Town, Beijing, PR China,. Tel.: +86 010 6177 2757.
E-mail address: ylqncepu@163.com (L. Yang).

http://dx.doi.org/10.1016/j.ijleo.2015.10.078
0030-4026/

and short-lived, so the phishing detection methods need to have strong real-time performance and intelligence [12]. Tsang et al. [13] proposed a method for detecting phishing pages by searching similar webpages through comparing the webpages by matching HTML source codes as well as computing the cosine similarity of detecting phishing pages. Most of the previous phishing webpage recognition techniques are aimed at single web page, the process of feature extraction depends on fewer and single page features, such as textual features, image features. These schemes are web-facing detection methods that cannot carry out comprehensive analyses on website. So, in addition to reducing the detecting efficiency, that are unable to achieve the goal of real-time detecting phishing website. SVM is a traditional classification algorithm, as it showed to be higher in classification accuracy, but the SVM algorithm delivers slower training speed when coping with large training sets, and that becomes a drawback in practical applications [14]. For the above reasons, in this paper, we based on the differences between the phishing websites and the imitated target websites, from the perspective of the Web topology structure, we use the BVM algorithm to detect and classify phishing site. In order to quantify the topological features, we first extract 12 statistic indices of websites as the topological features to complement the disadvantage of the detection based on single webpage features. Second, we have BVM instead of SVM, which then utilize BVM classifier to classify the features vectors, which are composed of topological features. As an independent scheme, the experimental results show that our method is efficient at detecting phishing sites, moreover, it can achieve higher accuracy of phishing detection and have faster training speed.

The rest of the paper is organized as follows: Section 2 describes in detail the theoretical basis of BVM classifier. Section 3 introduces the process of getting the raw data crawled by web crawler, obtaining the required features and forming the features vectors correspond to websites. Section 4 describes the datasets used in our experiments, and illustrates the experimental design and the performance of the proposed method. Conclusion and orientations for future works are discussed in section 5.

2. SVM classification via minimum enclosing ball

In recent years, many machine learning algorithms are developed for solving classification problems and have been widely applied for detection of phishing websites [15]. To construct high-performance phishing detection model, there are two aspects of work need to be done: firstly, getting a good performance classifier and obtaining high precision in classifying normal data and anomaly data; secondly, reducing the training time without impacting the classification accuracy. The SVM performed good capability in detecting phishing website. Solving quadratic programming (QP) problem in SVM becomes a decisive factor for classification problem, and that can be transformed into the process of solving minimum enclosing ball. In this paper, emphasis is given on the BVM algorithm, which is firstly applied the BVM algorithm to the phishing detection. Our method not only can reduce the complexity of the classification model, but also can improve the detection precision of the phishing website. The related concepts and algorithms of the BVM will be formulated as follows.

2.1. Quadratic programming problems in SVM

Support vector machine (SVM), to say it colloquially, is a binary classification model, which is defined as the maximum linear classifier in feature space margin. The learning strategy of the model is to maximize feature space margin, so the problem of solving maximum margin can be converted into that of a convex quadratic programming problem [16].

For a given set of $\boldsymbol{N}$ training samples $\{(x_i,y_i),i=1,2,\ldots,\boldsymbol{N}\}$, where $x_i \in R^m$ is M-dimensional feature vector of a training sample, and $y_i \in \{-1,+1\}$ is its known class label. Finding a hyperplane in the M-dimensional data space becomes the learning objective of the linear classifier, and the linear classifier can be formulated as $w^{\mathrm{T}}x+b=0$. To obtain the optimal hyperplane, the objective function of SVM can be formulated as the following problem:

$$\max \frac{1}{\|w\|} \tag{1}$$

$$\text{subject to}: y_i(w^T x_i + b) \geq 1-\xi_i,\ i=1,2,\ldots,n \qquad \xi \geq 0,\ i=1,2,\ldots,n \tag{2}$$

The above problem can be transformed into the equation as follows:

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \tag{3}$$

$$(4)\text{subject to}: y_i(w^T x_i + b) \geq 1-\xi_i,\ i=1,2,\ldots,n$$

As seen from (3), the objective function is quadratic, and the constraint is linear, so the SVM is a convex quadratic programming problem. All of this is to say that find the optimal hyperplane under some constraints and minimize the losses of creation time of the SVM classifier.

2.2. The equivalence of BVM and standard SVM

As mentioned above, the process of training a SVM contains a quadratic programming problem. If we define the training sample size is as n, the time complexity and space complexity of the SVM algorithm are $O(n^3)$ and $O(n^2)$, respectively. Therefore, the time complexity and space complexity will rise as the training set gets larger [13]. According to the principle of minimum enclosing ball (MEB), the optimized solution in SVM can be transformed into solving the minimum enclosing ball problem, which makes the classifier training process need a core subset of large-scale data set to be participated in training [17]. Consequently, improving the ability of dealing with large training set in detection of phishing websites. The equivalence relation between BVM and standard SVM can be proved below:

Given a dataset $S=\{x_i\}_{i=1}^{m}$, $x \in R^m$, k is the kernel function which depends on nonlinear mapping ϕ, that is $k(x_i,x_j)=\langle\phi(x_i),\phi(x_j)\rangle$. The process of solving the minimum enclosing ball of S can be formulated as follows:

$$B(c^*,R^*)=\arg\min_{c,R} R \tag{5}$$

$$(6)\text{subject to}: A\left\|c-\phi(x_i)\right\|^2 < R^2\ \forall i$$

The above problem can be transformed its dual problem that is to maximize the equation as follows:

$$\max_{\alpha_i} \sum_{i=1}^{m}\alpha_i k(x_i,x_i) - \sum_{i,j=1}^{m}\alpha_i\alpha_j k(x_i,x_j) \tag{7}$$

subject to:

$$\sum_{i=1}^{m}\alpha_i = 1,\ \alpha_i \geq 0,\ i=1,2,\ldots,m \tag{8}$$

The core function of the above equation can be formulated as follows:

$$k(x,x)=\tilde{k} \tag{9}$$