IJCAT - International Journal of Computing and Technology, Volume 3, Issue 5, May 2016
ISSN : 2348 - 6090
www.IJCAT.org

Support Vector Machine Based Malware and Phishing Website Detection

[1] Rashmi Karnik, [2] Dr. Gayatri M. Bhandari

[1, 2] Department of Computer Engg., JSPM's Bhivarabai Sawant Institute of Technology & Research
Savitribai Phule University, Pune.

Abstract - The use of Internet leads to various security threats. It includes spamming, phishing or malware. The phishing attack retrieves the sensitive information like bank account number or email password etc. Most of the phishing attack use malicious URL. The Malicious URL will be displayed to the user like a legitimate URL. Malware is widely used to disrupt computer operation, gain access to users' computer systems or gather sensitive information. Nowadays, malware is a serious threat of the Internet. Detecting malicious URLs is an essential task in network security intelligence. In this paper we categories phishing and malware URLs using Support Vector Machine (SVM). The Support Vector Machine (SVM) is a widely used kernel-based method for binary classification. SVM is theoretically well founded and has been already applied to many practical problems. Our method uses a variety of discriminative features including textual properties, link structures, webpage contents, DNS information, and network traffic. It shows that our proposed method is good at detecting phishing and malware sites, correctly labeling approximately 95% of phishing and malware sites. We achieve high performance, including high level of true positive, true negative, sensitivity, precision, F-measure and overall accuracy compared with other approaches. So we can say SVM is a robust and efficient method that can be successfully used for classification of normal or phishing website.

***Keywords* - Kernel based Approach, Malware, Phishing Support Vector Machine.**

1. Introduction

Phishing is a form of identity theft that occurs when a malicious Web site impersonates a legitimate one in order to acquire sensitive information such as passwords, account details, or credit card numbers. Though there are several anti-phishing software and techniques for detecting potential phishing attempts in emails and detecting phishing contents on websites, phishes come up with new and hybrid techniques to circumvent the available software and techniques.

To prevent a user from browsing phishing sites, there are two distinct approaches. One is URL filtering. It detects phishing sites by comparing the URL of a site a user visits with a URL blacklist composed of the URLs of phishing sites. However, it is difficult to build a perfect blacklist due to the rapid increase of phishing sites. When the URL of the site is not registered on the URL white-list, the site will be marked as a phishing site. A URL white-list is composed of URLs of legitimate sites and is able to detect phishing sites because URLs of phishing sites cannot be registered on the white-list. However, it is extremely difficult to register large numerous numbers of legitimate sites. With the increasing severity of this issue, many efforts have been devoted to apply machine learning methods to phishing detection. One of the most common machine learning techniques for phishing classification is to use a list of key features to represent an email and apply a learning algorithm to classify an email to phishing or ham based on the selected features.

Blacklisting is the most common anti-phishing technique used by modern web browsers. However, study [1] shows that centralized, blacklist-based protection alone is not adequate enough to protect end users from new and emerging zero-day phishing webpages that appear in the thousands and quickly disappear every day.

2. Literature Survey

2.1 Non-Content Based Approaches

Non-content based approaches include URL and host information based classification of phishing sites, blacklisting and white-listing methods. In URL based

schemes, URLs are classified based on both lexical and host features. Lexical features describe lexical patterns of malicious URLs. These include features such as length of the URL, the number of dots, special characters it contains. Host features of the URL include properties of IP address, the owner of the site, DNS properties such as TTL, and geographical location [5]. Using these features, a matrix is built and run through multiple classification algorithms. In real-time processing trials, this approach has success rates between 95-99%.

PhishNet [7] processes blacklisted URLs (parents) and produces multiple variations of the same URL (children) via 5 different URL variation heuristics, which are Replace Top Level Domains, Directory structure similarity, IP address equivalence, Query string substitution and Brand name equivalence.

White-listing approaches seek to detect known good sites, but a user must remember to check the interface every time he visits any site. Automated Individual White-List (AIWL) [8] maintains a whitelist of features describing trusted Login User Interfaces (LUIs) where the user submitted his/her credentials.

2.2 Content Based Approaches

In content based approach, phishing attacks are detected by examining site contents. Features used in this approach include spelling errors, source of the images, links, password fields, embedded links, etc. along with URL and host based features. SpoofGuard [1] and CANTINA [9] are two such approaches. SpoofGuard detects HTTP(S)-based phishing attempts as a web browser toolbar, by weighting certain anomalies found in the HTML content against a defined threshold value. It also uses history, such as whether the user has visited this domain before and whether the referring page was from an email site such as Hotmail or Yahoo! Mail. CANTINA examines the content of a web page to determine whether it is legitimate or not, in contrast to other approaches that look at surface characteristics of a web page [4].

2.3 Support Vector Machine

Support vector machines are an example of supervised learning algorithms which belong to both the regression and classification categories of machine learning algorithms. SVMs is a collection of machine learning algorithms that can be used to recognize patterns in given data. Given a set of training data it would like to classify. A classification task usually involves separating data into training and testing sets. The goal of SVM is to produce a model (based on the training data) which predicts the target values of the test data [12]. SVM method does not suffer the limitations of data dimensionality and limited samples. Several recent studies have reported that the SVM (support vector machines) generally are capable of delivering higher performance in terms of classification accuracy than the other data classification algorithms. It has been employed in a wide range of real world problems such as text categorization, hand-written, digit recognition, tone recognition, image classification and object detection, micro-array gene expression data analysis, data classification [2]. SVM acts as a machine learning based system for the detection of malware [3].

3. Related Work

3.1 Existing Systems

Most of the malicious detection methods use any one of the following methods: Non Content based and Content based System. The non content based approach, extract the URL's host information and DNS information. The content based approach, extract the content of the URL's that contains source of image, inner links and outer links of the web page, and tags of the pages. It works with following processes:

1) Malware Feature Extraction: The malware feature extraction can be categories into static, dynamic and hybrid. Dynamic analysis techniques observe the execution of the malware to derive features. Static analysis techniques analyses the malware without running it. The analysis target can be binary or source code. Hybrid analysis is an approach that combines static and dynamic analysis to achieve the both benefits.

2) Malware Categorization: Different classification approach including association classifiers, support vector machines, and Naive Bayes have been applied in malware detection. Malware families detects by HOLMES [9][4] combines frequent sub graph mining and concept analysis to synthesize selective specifications. For building classification model it require to frame a large number of training samples.

3) Phishing Website Detection: Many detection methods like say support vector machine or say naïve bayes for detection of phishing websites [2]. But the reality is that today there exist only few methods which efficiently detect phishing website detection using clustering approach. It applies clustering algorithm to decide if any cluster exists around given web page. In case any webpage finder then is treated as phishing webpage or either as genuine page.

4) Ensemble Clusters: Cluster ensemble is used to aggregate the clustering solutions that are generated by both hierarchical and partitional clustering methods. The clusters are formed as group of components sharing some common properties and thus dissimilar components are placed at different clusters.

3.2 Domain Knowledge

Machine learning is a mature and well-recognized research area of computer science, mainly concerned with the discovery of models, patterns, and other regularities in data. Machine learning approaches can be roughly categorized into two different groups:

- ***Symbolic approaches.*** Inductive learning of symbolic descriptions, such as rules decision trees or logical representations.
- ***Statistical approaches.*** Statistical or pattern-recognition methods, including k-nearest neighbor or instance-based learning, Bayesian classifiers, neural network learning, and support vector machines.

3.3 Problem Definition

Phishing is a type of computer attack that communicates socially engineered messages to humans via electronic communication channels in order to persuade them to perform certain actions for the attacker's benefit. Let u_1, u_2, …u_n be a list of URLs where each u_i has the list of features (f_1, f_2,…,f_m), an machine learning approach is applied to detect an u_i is legitimate, phishing, or malware URL.

3.4 Proposed System

We are using Support vector machine (SVM) algorithm for categorizing phishing and malware sites. SVM is a supervised machine learning algorithm. It efficiently classified the malicious URL. It contains two phases Training Phase and Testing Phase.

In the training phase, using SVM kernel function (Linear/Gaussian, Polynomial and Sigmoid) the SVM model is generated. Based on the generated model, the test data is classified.

In the testing phase, all the features are extracted from test URL. The extracted features are classified based on the training data set.

4. System Architecture

Fig 1. shows the combined system architecture of existing and proposed system. Its components are briefly described below:

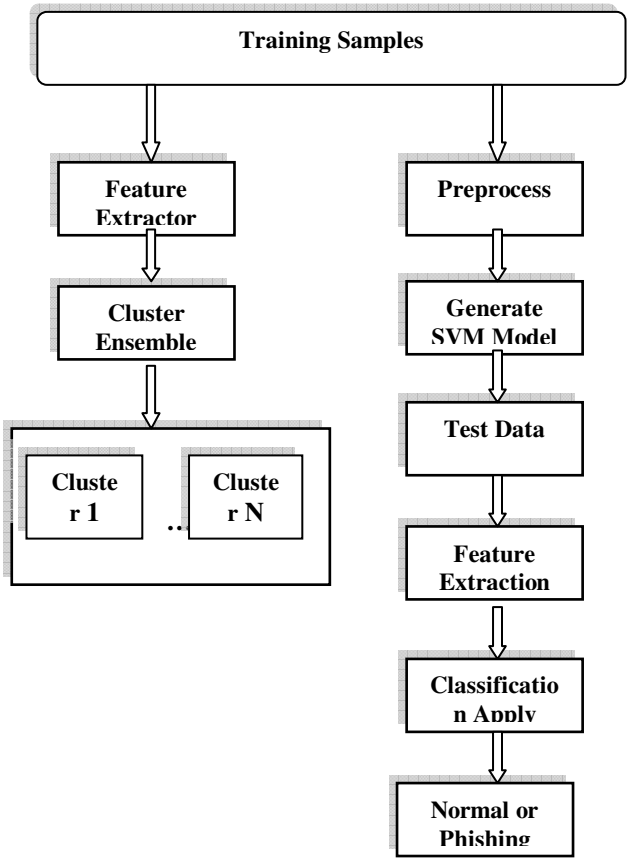

Fig 1. System Architecture

1) Feature extractor: Feature extractor extract the terms from the webpages of the collected phishing websites, and then transforms the data into term-frequency feature vectors. These vectors are stored in the database.

2) Cluster Ensemble: Cluster ensemble is used to combine different base clustering's. Base clustering solutions are generated by applying different clustering algorithms that are based on the feature representations. The HC algorithm and KM partitioned approach are applied on the term-frequency vectors. This helps in formation of clusters.

3) Preprocess: The training data set is collected from the internet. The collected data set is re-processed. Preprocessing is followed by removing the record,

which contains any missing values. This is called training phase of SVM.

4) Generate SVM Model: The train data set contains set of URL with number of features. Initially the data set is trained using SVM algorithm by using the kernel functions such as linear, polynomial or sigmoid, which generates SVM model.

5) Feature Extraction: Given a URL(test data) the feature extractor extracts all the features of URL based on the above SVM model and categorizes the web site to be normal or phishing.

5. Design Process

5.1 Base Clustering

A cluster is a collection of phishing websites or malicious files that share some common traits between them and are "dissimilar" to the phishing websites or malware samples belonging to other clusters. Hierarchical and partitioning clustering are two common types of clustering methods, and each of them has its own traits [2].

1) Hierarchical Clustering Algorithm: Hierarchical algorithm starts as frame with N singleton clusters and then successfully merges two nearest clusters until only one remains. This technique is suitable for both phishing and malware detection or categorization.

Hierarchical Clustering Algorithm

Input: The Data set D
Output: The best K and data clusters

Set each data point as a singleton cluster;
For K ← N-1 to 1 do
1. Merge two closest clusters C1 and C2 into new cluster C with C1+C2 elements;
2. Calculate the similarity from C to all other clusters and update the similarity matrix;
3. Calculate the validity index;
4. Compare and keep the best K and corresponding clusters until now;

End
Return the best K and corresponding clusters.

2) K-Medoids Clustering Approach: This approach assigns a set of data points into clustering an iterative relocation technique [2]. A cluster is represented by one of its real data point (called medoids) or by the mean of its data points (called centroid) in KM and K-means methods, respectively. They are very simple, but effective and widely used in many scientific and industrial applications.

K-Medoids Clustering Algorithm

Input: N points in d-dimensional space, number of clusters k
Output: k clusters
Randomly choose k cluster Medoids
Repeat
1. Assign each point to the nearest cluster;
2. Update the cluster medoid by the calculation of validity index;

Until the medoids do not change;

5.2 Support Vector Machine

The SVM algorithm is applied to categories the phishing and malware site. SVM is machine learning algorithm. SVMs classify data by determining a set of support vectors, which are members of the set of training inputs that outline a hyper plane in the feature space. SVM contains three functions linear, polynomial and sigmoid. User can select any one of the function to classify the data. We will be dealing with Linear/RBF(Radial basis function) or Sigmoid kernel method.

SVM Algorithm

Input: Train Data Set - Train, Test Data Set – Test
Output: Web site Categorization Result Normal, Phishing or Malware
1. Read Train Data Set
2. Apply SVM algorithm
3. Generate SVM Model for kernel function
4. Read Test Data Set
5. **For** each URL in Test Data
6. Extract all the features
7. Apply SVM algorithm
8. Return Result of Test Data
9. **End**

5.3 Mathematical Model for proposed System

Input: Train Data Set - Train, Test Data Set – Test
Process: The following parameter is needed to generate the SVM model:

N: - Total number of features
K:- Kernel Function
X:- Feature Vector

Kernel is defined as a function that accepts two vectors x_i

and x_j as inputs and produces an output which is defined as the inner product of their images $\Phi(x_i)$ and $\Phi(x_j)$

$$K(x_1, x_2) = \Phi(x_1)^T \Phi(x_2)$$

The main idea here is to generate a learning algorithm that operates in kernel space, which is generated by substituting the values of all inner products from the original space into the newly formed kernel space:

$$f(x) = \phi(x)^T w + b = \sum_{j=1}^{n} \alpha_j y_j K(x, x_j) = b$$

The parameter b can be found from any support vectors x_i

$$b = y_i - \phi(x_i)^T w = y_i - \sum_{j=1}^{n} \alpha_j y_j (\phi(x_i)^T \phi(x_j)) = y_i - \sum_{j=1}^{n} \alpha_j y_j K(x_i, x_j)$$

The four basic kernels are given as follows [12]:

- Linear:

$$K(x_i, x_j) = x_i^T x_j$$

- Polynomial:

$$K(x_i, x_j) = \gamma(x_i^T x_j + r)^d, \gamma > 0$$

- Radial basis function (RBF):

$$K(x_i, x_j) = \exp(-\gamma(\| x_i - x_j \|^2), \gamma > 0$$

- Sigmoid:

$$K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r), \gamma > 0$$

Here γ, r, and d are kernel parameters.

SVM model is generated for the training data set. Extract all the features of test data with the help of above kernel methods. Apply SVM algorithm to categories web site. For each URL in test data is classified as Normal of Phishing Site.

Output: Web site Categorization Result Normal, Phishing or Malware

5.4 Modules

The system consists of following modules:

- Preprocessing
- Feature Extraction
- Classification

Preprocessing: The training data set is collected from the Internet. The collected data set is preprocessed. Remove the record, which contains any missing values. Check all the records contains class label (Normal or Phishing).

Feature Extraction:
In this module the following features are extracted from the URL

- IP address - IP address in the domain name of the URL
- Long URL – Length of the URL
- URL's having @ symbol – URL contains @symbol
- Prefix and suffix – domain part has '- '
- Sub-domain (dots) - dots in domain name
- Misuse/fake of HTTPs protocol – Not using https protocol
- Request URL - objects are loaded from a domain other than the URL
- Server form handler – The server transferred data to another domain.
- URL of anchor – No of anchor tag in URL page
- Abnormal URL – No host name in URL
- Using pop-up window -Usually authenticated sites do not ask users to submit their credentials via a popup window
- Redirect page – Redirect to suspicious page
- DNS record – Empty DNS record
- Hiding the links - change of status bar onMouseOver
- Website traffic – Determine traffic rate
- Age of domain- Presence of web site

Classification: In this module the test data can be classified Normal or Phishing using SVM algorithm. Based on generated SVM model the test data is classified. For each URL extract the features and classify the URL using SVM.

5.5 System Requirements

Hardware requirements:

Processor	Any Processor above 500 MHz
Ram	2 GB
Hard Disk	10 GB
Compact Disk	650 Mb
Input device	Standard Keyboard and Mouse

Software requirements:

Operating System	Windows XP or Windows 7,8
Technology	Net Beans 8.0 Jdk1.7
Database	MySQL

6. Results

Fig 2. shows the comparison between the existing and proposed system. The time required by using SVM to make categorization has been reduced by few milliseconds as compared to existing system that uses clustering algorithm to make categorization of phishing or normal website. For testing any website existing system take approximately 98ms while proposed system takes 93ms.

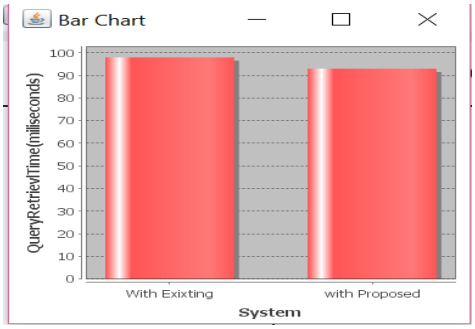

Fig. 2. Existing vs proposed results in Milliseconds

7. Conclusion

Detecting the malicious URL is one of the crucial problems in internet. This paper investigates the problem of web site categorization i.e., Normal or Phishing. This paper presents the supervised machine learning approach SVM is used to categories phishing and malware sites. This paper extracts various numbers of features from the URL. The Support vector machine algorithm achieved high classification accuracy for analyzing similar data parts to those of rule-based heuristic techniques. Our proposed method is good at detecting phishing and malware sites, correctly labeling approximately 95% of phishing and malware sites.

References

[1] Ram B. Basnet, Andrew H. Sung, Quingzhong Liu, "Learning To Detect Phishing URLs", IJRET: International Journal of Research in Engineering and Technology, Volume: 03 Issue: 06 | Jun-2014.

[2] Usha Narra, Corrado Aaron Visaggio, Mark Stamp, Thomas H. Austin, "Clustering versus SVM for malware detection", Springer, Journal of Computer Virology and Hacking Techniques 10/2015

[3] Anjali B. Sayamber ,Arati M. Dixit , "Malicious URL Detection and Identification", *International Journal of Computer Applications (0975 – 8887) Volume 99 – No.17, August 2014.*

[4] Michal Kruczkowski; Ewa Niewiadomska Szynkiewicz, "Support Vector Machine for Malware Analysis and Cla ssification" Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences

[5] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Identifying Suspicious URLs: An Application of Large-scale Online Learning," in ICML '09: Proceedings of the International Conference on Machine Learning, 2009, pp. 681–688.

[6] C. Whittaker, B. Ryner, and M. Nazif, "Large-scale automatic classification of phishing pages," in NDSS '10, 2010.

[7] P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, "Phishnet: predictive blacklisting to detect phishing attacks," in INFOCOM'10: Proceedings of the 29th conference on Information communications. Piscataway, NJ, USA: IEEE Press, 2010, pp. 346–350.

[8] Y. Cao, W. Han, and Y. Le, "Anti-phishing based on automated individual white-list," in DIM '08: Proceedings of the 4th ACM workshop on Digital identity management. New York, NY, USA: ACM, 2008, pp. 51–60.

[9] Y. Zhang, J. Hong, and L. Cranor, "Cantina: A Content-based Approach to Detecting Phishing Web sites," in proceedings of the International World Wide Web Conference (WWW), 2007.

[10] M. Fredrikson, S. Jha, M. Christodorescu, R. Sailer, and X. Yan, "Synthesizing near-optimal malware specifications from suspicious behaviors," in *Proc. IEEE Symp. Secur. Priv.,* Washington, DC IEEE Computer Society, May 2010, pp. 45–60

[11] A. Y. Fu, L. Wenyin, and X. Deng, "Detecting phishing web pages with visual similarity assessment based on earth mover's distance (emd)," IEEE Trans. Dependable Secur. Comput., vol. 3, no. 4, pp. 301–311,2006.

[12] A Practical Guide to Support Vector Classification Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin Department of Computer Science National Taiwan University, Taipei 106, Taiwan http://www.csie.ntu.edu.tw/~cjlin Initial version: 2003 Last updated: April 15, 2010.

[13] M. Hara, A. Yamada, and Y. Miyake, "Visual similarity-based phishing detection without victim site information," in IEEE Symposium on Computational Intelligence in Cyber Security, 2009. CICS '09, 2009, pp. 30 – 36

[14] Michael Atighetchi, Partha Pal "Attribute-based prevention of phishing attacks" Eighth IEEE international symposium on network computing and application, 2009.

[15] Matthew Dunlop, Stephen Groat, and David Shelly" GoldPhish: Using Images for Content-Based Phishing Analysis", in proceedings of internet monitoring and protection(ICIMP),fifth international conference, Barcelona, Pages 123-128, 201