IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

Difference Convex Programming based TSVM to Detect Phishing Pages

K. Selvan[1], Dr. M. Vanitha[2]

Research Scholar, Department of Computer Application, JJ College of Arts and Science, Pudukkottai, India [1]

Research Advisor, Department of Computer Science, Alagappa University, Karaikudi, India [2]

Abstract: Phishing sites are duplicated created to trick people for their personal information, in order to simulate the real site's pages. Because they rarely detect tactical adaptability costs and identify phishing site it is a complex and dynamic problems. This work proposes a new method to detect phishing sites using Convex Programming based Transductive Support Vector Machine (CTSVM). TSVM is an independent method to detect attack, and does not change the behavior of users proposed a new way of phishing sites. Image feature extraction, and sensitive information on the page, it can completely reflect the nature of the site website. Then, phishing pages are classified by the algorithm CTSVM. Experiment result shows that the method performed well, improving the accuracy and precision are more fully. The result is to use CTSVM detect phishing sites to improve performance, such as classification TSVM should be more flexible, which makes learning CTSVM, in order to more efficiently.

Keywords: Classification, convex programming, feature selection, phishing websites, transductive support vector machine.

I. INTRODUCTION

Phishing can be defined as the use of social engineering techniques in criminal activities. The most common example of phishing is an email requesting access, account information / credit card, e-commerce website (eBay, Amazon, etc.) and online banking. It is the kind of merger of Web technology and social engineering. The most popular phishing carried out using phishing web sites. Phishing sites mimic the shape of some of the real company website. This is the kind of information security attacks. Phishing, vishing and smishing use both methods allow the thief to use the latest technology for personal customer account information for phishing purposes. In general, phishing detection technology can be divided into several categories, because the detection solution deployment.

In order to prevent a user's browser phishing site, there are two different methods. One of them is a filter in the URL. A URL site user's visit phishing sites compare consisting URL blacklists URLs detect phishing sites. However, it is difficult to establish a perfect blacklist because of the rapid increase in phishing sites. When the address of the website is not registered in the URL white list, the site will be flagged as phishing sites. A URL white list includes a legitimate Web site URL and capable of detecting phishing sites fishing, because the site's URL cannot be registered in the white list.

However, it is very difficult to record a lot of a lot of legitimate sites. Worsening of the problem, they have devoted much effort to detect phishing apply machine learning methods. The most common technique in which a machine learning classifier Phishing is the use of a list of key features to represent the basis of an e-mail application and the learning algorithm, an e-mail, phishing or ham classification selected function.

Anti-phishing blacklist modern web browsers most commonly used technique. However, this study [1] shows that, based on a single centralized blacklist protection is inadequate protection of new and emerging phishing site now thousands of end users zero sunrise soon disappear every day.

II. RELATED WORK

L. Wu, X. Du and J. Wu explore a convex approximation of the TSVM relaxation optimization problem. They present that takes the form of a semi-definite program. While this program can be solved in polynomial time, it becomes too inefficient for test sets with more than 100 examples. However, assuming a low-rank structure of the test labels derived from a spectral decomposition technique [2].

W. Chu, B.B. Zhu et al describes the difficulty in solving the TSVM optimization problem has led to much interest in other formulations of transductive learning algorithms. The goal is to exploit the same type of relationship between the geometry of the test examples — or unlabeled examples more generally — and their labels, but that has computationally more convenient properties. Graph partitioning approaches based on st-min-cuts and spectral graph partitioning explicitly or implicitly pursued this goal [2].

H. Kordestani and M. Shajari describes in the semi-supervised learning literature, methods have been

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

proposed from different perspectives, including margin-based classification, the EM method, graph-based method, and information regularization. The central topic this article concerns is the generalization performance of transductive support vector machine (TSVM), which remains mysterious, particularly its "alleged" unstable performance in empirical studies [3].

G. Liu, B. Qiu argues that in principle TSVM performs no worse than its supervised counterpart SVM after tuning. Key to it is tuning, which has been commonly ignored in the literature. Tuning guards against potential unstable performance by tuning regularizes towards labeled data. Furthermore, develop a statistical learning theory to demonstrate this aspect with regard to TSVM's generalization ability. To treat the implementation issue, develop a non convex minimization routine based on recent advances in global optimization, particularly difference convex (DC) programming. Numerical analysis indicates that the proposed routine delivers a better solution, and confirms that TSVM performs no worse than SVM [7].

III. TYPE OF PHISHING

A. Session Hijacking
It's also known as cookie session hijacking kidnapping. It is the user session security attacks. IP spoofing and middle attack is the most common method of session hijacking. HTTP communication needs different TCP connection and Web server to use the pool to recognize these connections. Usually these chips are a series of variables such as URL or HTTP header or other parts of the body.

B. Key loggers
Keylogger is a monitoring by people with the keyboard behavior monitoring software. Written in the keyboard information is recorded, the data is stored in a specific log file, and send it to a higher level. For example, in a commercial environment, it is mainly used to monitor employees' activities to ensure that they do not use a computer and only commercial use of other harmful activities. Sometimes, however, the log file is passed to an unknown third party. This may cause your information to unknown parties.

C. System reconfiguration attacks
Malicious nature of the attack system reconfiguration is changing the configuration of the user's PC. For example, you can change the file's user favorite, favorite files often contain a user visited URL. So modify the file may lead to malicious attacks.

D. Malware-based phishing
Based malware, phishing, spoofing or message alert users to upgrade. From the computer's alert notifications or update malicious software users. These updates information contained in the message, such as "computer has a virus, you need virus protection." Therefore, these messages may insist that users perform some unnecessary actions that resulted in malicious attacker to obtain similar credit card personal information.

E. Data Theft
Data theft is known as the company's confidential data theft and abuses it. Who is the person stealing the data can work in the same company or he / she may have left the company.

F. Search engine phishing
Fake website search engine created by phishing attackers appear. Here, the attacker waits for foreign customers based on sales or registration process, your personal information in places such as order.F. Search engine phishing

G. Pharming
Pharming refers to malicious practice is redirected to one of various types of phishing attacks performed on fraudulent websites without their consent DNS server by creating false information and users.

H. Web Trojans
The reason is the Trojan site login create false pop up on the screen to collect user credentials.

I. Deceptive Phishing
The term "phishing" generally refers to personal information stolen via instant messaging. But today, a fraudulent e-mail messages to be used as common methods of phishing. In other words it refers to fraudulent phishing websites created a different impression of the original, so gullible false personal information in response to the link and pass it.

J. Host File Poisoning
When users try to navigate to the Web site, users can use the DNS server or a local file to determine either the IP address of the computer is called the host. Host files poisoning, hackers can fake site from a legitimate Web site, they can steal confidential information from the user's user redirection.

IV. METHODOLOGY

1. Design and implement anti-phishing plug-in web browser and assessment, provide basic information and block phishing phishing sites, so users can protect the Semantic Web attacks. This paper shows that computer users can be trained, and feed confidential information on the site, in order to make better decisions.
2. Yield with different people put forward anti-phishing system discussed from different fields. Anti-phishing system checked and compared with other options, Available on the Internet. Our analysis led to many findings and recommendations to improve phishing countermeasures.
3. The proposed system of false positives, false negative, true-positive and true-negative cases details. This means that the system model to differentiate based on their behavior and the different sites. We have studied the

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

effectiveness of web browser used by the popular fishing tools. It has been found, no tools provide accurate and timely results, so using the proposed system can be advantageous for the Internet users. It has been found that the black list to protect the user during the initial is invalid. The tool uses heuristics to supplement trapped significantly more phishing sites blacklist.

4. The system is designed to store information and legitimate website blacklist, and shows how sites are declared as black or legalization list. By using this updated information, the Internet user can know the website. By looking at the screenshots, the user can determine the type of site, if it looks like fishing; the user can check the tool to send information on the website. Check the page, the tool system gives users instant results.

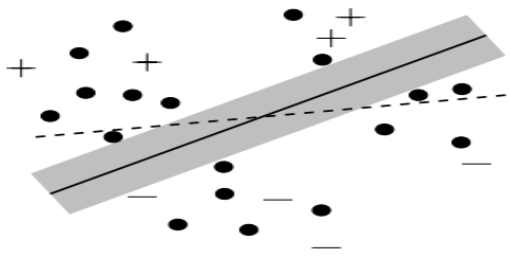
V. ALGORITHM USED

A. Transductive support vector machines (TSVMs)

Direct support vector machine (TSVMs) assume specific geometric relationship between X = (X1, ..., XN) and P (Y1, ..., YN). They built on the upper edge of the structure H hyperplane {X: W • X + B = 0} the full sample, X = (X1, X2, ..., xn) mapping, including training and test vector. In X is X margin hyperplane in the example, the minimum distance between the closest carriers

$$\min_{i\in[1..n]}\left[\frac{\mathbf{y}_i}{\|\mathbf{w}\|}(\mathbf{w}\cdot\mathbf{x}_i+b)\right]$$

Hρ structure contains elements with hyperplane H (x) = {W + symbol width x •} to achieve at least ρ ρ dependence marker Hρ description of all ten pairs margin X in Figure 1, intuitively, based on the structure structure of the balance of priorities following cluster clusters across borders more tags cut mark.



B. Difference convex programming

DC is an exploded key programming function of the difference in the cost of two convex function of a sequence of high approximate sequence convergence solutions to obtain a stationary point, probably based on the minimum -ε global cost function. This technique is called DC algorithm and ψ learning and plenty of semi-supervised learning applications for the big problems has been used.

$$C_1\sum_{i=1}^{n_l}L(y_i f(x_i))+C_2\sum_{j=n_l+1}^{n}L(|f(x_j)|)+J(f).$$

C. Algorithm 1: (TSVM^{DCA})

Step 1. (Initialization) Set initial value f(0) as the solution of SVM with labeled data alone, and an precision tolerance level $\epsilon > 0$.

Step 2. (Iteration) At iteration k+1, solve yielding solution f(k+1). The dual problem can be solved yielding the solution, as described in

Step 3. (Stopping rule) Terminate when $|s(f^{(k+1)}) - s(f^{(k)})| \leq$ ϙ. Then the estimate f ˆ is the best solution among $f^{(k)}$; k = 0, 1, ·

A good initial value nevertheless enhances the chance of Algorithm 1 to locate the global minima. Our numerical experience suggests that SVM is an acceptable choice.

VI. IMPLEMENTATION

A. Evaluation Metrics

By comparing the actual prediction classification category, we can calculate (TN each attribute correctly classified) the number of true negatives and false negative (FN, phishing attribute misclassified), true positive (TP, correct classification of each phishing) and false positive (FP, attributes are mistaken for phishing detection). In order to evaluate the performance of the sounder classification accuracy (ACCU), weighted accuracy (WACC) by the following formula:

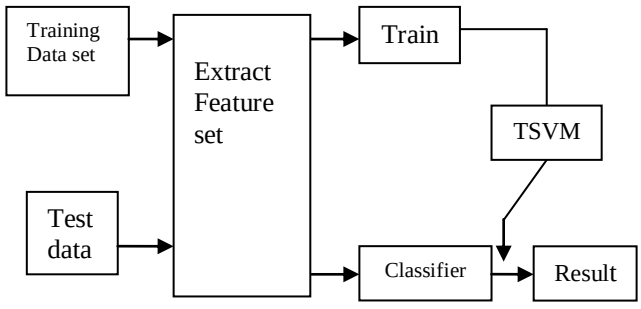
$$\text{Accu} = \frac{\text{TN}+\text{TP}}{\text{TN}+\text{FP}+\text{TP}+\text{FN}} \quad (1)$$

$$w_{acc}(\lambda) = \frac{\lambda.\text{TN}+\text{TP}}{\lambda.(\text{TN}+\text{FP})+\text{TP}+\text{FN}} \quad (2)$$

B. Dataset

The data set using the learning library phishing sites from UCI machine. This set of data collected mainly Phishtank data files, MillerSmiles file, Google search operators. None 30 attributes implementation and the results gathered by the almost 2456 instances and operations.

C. Results



The original data of intrusion detection, having very high dimension and including dozens of properties, will classify very slowly or even unable to go on if these data are directly applied to a number of classification algorithms. Furthermore, not all the features of the original data have positive impact on test results, and some features may even

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

affect the test results. Therefore, it is necessary to preprocess the data before they are sent to the parser to filter the characteristics having little influence or negative impact on test results and try to keep the key features of the data. We employ the Manifold learning to solve this problem, which is used as a data reduction method. Manifold learning can map high dimensional data points to a lower dimension, so we can achieve our goal of reducing the huge dataset. After using Manifold learning as our pre-processing method, we use one class Ball Vector Machine and TSVM to classify the data which is extracted from the original data.

TABLE 1: COMPOSITION OF TRAINING DATASET

	Normal	DoS	R2L	Probing
Training Dataset	52	126	80	222
Test Data	12	22	50	50

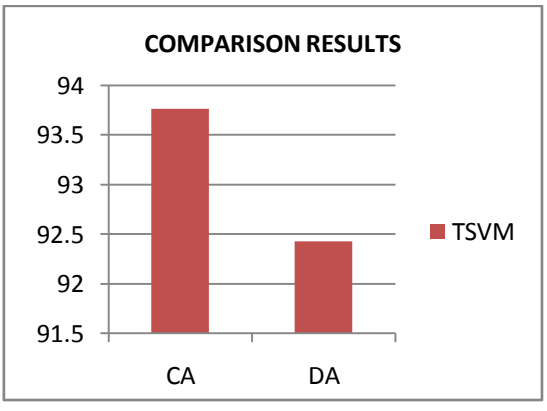
DoS - Denial of Service
R2L - Unauthorized Access from a Remote Machine, e.g. guessing password;
Probing - Surveillance and other probing, e.g., port scanning.
The purpose of the experiment is to evaluate the performance of proposed intrusion detection method. There are four steps in our experiment.

Step1. Map the training data set into a d-dimensional space using improved manifold learning algorithm.
Step2. Train TSVM using d-dimensional vectors of training data set.
Step3. Map the testing data set into a d-dimensional space using improved manifold learning algorithm.
Step4. Complete the testing task by using the d-dimensional vectors of test data set and the trained TSVM classifier.

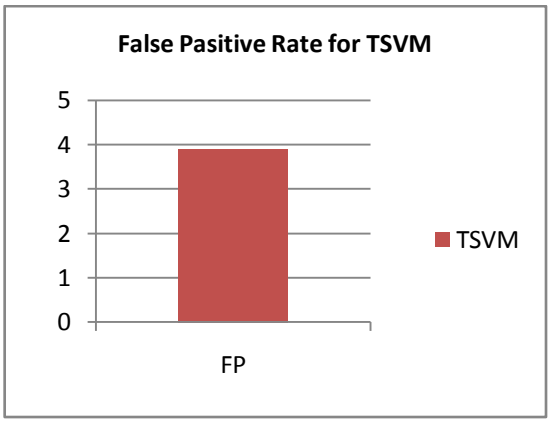
The results of experiments are shown in Table 2.

TABLE 2: RESULTS

	CA	DR	FP	Time (in Sec)
TSVM	93.76	92.43	3.9	0.11



CA = (number of correctly classified samples)/ (the total number samples);
DR = (number of detected intrusion samples)/ (the total number of intrusion samples);
FP = (number of normal samples that were detected as intrusion)/ (total number of normal samples).



VI. CONCLUSION AND FUTURE WORK

Phishing website detection is one of the key issues on the Internet. This paper describes the approach supervised CTSVM learning machine is used for phishing and malware sites category. In this work features several numbers phishing website is extracted. The convex programming based transductive support vector machine algorithm to achieve the technical part of the analysis based on heuristic rules similar to those data, high classification accuracy. Our proposed method is good at detecting phishing sites and malwares, phishing sites, and correctly labeled malware about 94%.

REFERENCES

[1] Anti-Phishing Working Group (APWG), "Phishing activity trends report second quarter 2014," http://docs.apwg.org/reports/apwg_trends_report_q2_2014.pdf, 2014, accessed Oct 2014.
[2] L. Wu, X. Du and J. Wu. "MobiFish: a lightweight anti-phishing scheme for mobile phones", in 23rd International Conference on Computer Communication and Networks, 2014, IEEE, pp. 1-8.
[3] W. Chu, B.B. Zhu, F. Xue, X. Guan and Z. Cai, "Protect sensitive sites from phishing attacks using features extractable from inaccessible phishing URLs," International Conference on Communications, 2013, IEEE, pp. 1990-1994.
[4] H. Kordestani and M. Shajari. "An entice resistant automatic phishing detection," in 5th Conference on Information and Knowledge Technology, 2013, IEEE, pp. 134-139.
[5] APWG 2nd Quarter 2014 Phishing Activity Trends Report from www.antiphishing.org
[6] Phishing website list from http://www.phishtank.com/, November 2013.
[7] G. Liu, B. Qiu and L. Wenyin. "Automatic detection of phishing target from phishing webpage," in 20th International Conference on Pattern Recognition, 2010, IEEE, pp. 4153-4156.
[8] J. Crain, L. Opyrchal and A. Prakash. "Fighting phishing with trusted email," in 10th International Conference on Availability, Reliability and Security, 2010, IEEE, pp. 462-467.
[9] J. Hiou and Q. Yang. "Defense against mobile phishing attack," in EECS 588 Project, 2012.
[10] [J. Hong, "The state of phishing attacks," Communications of the ACM, vol. 55, no. 1, pp. 74-81, 2012.