Kernel K-Means Clustering for Phishing Website and Malware Categorization

Kanti Sahu
Department of Information Technology
SATI College
Vidisha, M.P. (India)

S K. Shrivastava, Ph.D
Department of Information Technology
SATI College
Vidisha, M.P. (India)

ABSTRACT

In these days there are two famous internet attacks these are malware and phishing. Malware stands for malicious software. It is designed to damage computer system without knowledge of the user. Phishing website is comparatively new internet crime to malware attack. Phishing is a form of online fraud such as social engineering schemes by sending e-mails, sudden message or online advertising attract users to phishing website that pretend to be trustworthy website in order to trick individuals sensitive information for illustration- financial accounts, password and personal identification numbers, which is used for profit. Malware and Phishing website is share same properties, firstly increasing at a rate of thousands per day and secondly phishing webpage represented by the term frequencies of the website content share comparable characteristic of malware samples represented through instruction frequencies of the program executable code. Past few years many techniques have been develop to detect malware and phishing website. In these techniques firstly extract feature from phishing website or malware and then categorize them into group. In this paper, we proposed Kernel k-means clustering to categorize malware and phishing website. Kernel k-means is advance version of the k-means algorithm. In which vectors are mapped from vector space to a higher dimensional feature space through kernel function and then k-means is applied in feature space. Thus kernel k-means avoids the separable clusters in vector space and improves the accuracy of phishing website and malware categorization.

General Terms

Clustering, Internet attack.

Keywords

Malware, Phishing website, Kernel k-means clustering algorithm.

1. INTRODUCTION

Malware Categorization - Malware includes computer viruses, worms, Trojan horse, spyware, adware, scare ware, ransom ware and other malicious programs. Malware is a combined term for any malicious software which enters system without authorization of the owner using system. Malware which is malicious software is a big threat in today's computing world. It continues to produce in huge amount. As more as organization try to address the obscurity, the number of sites distributing the malware is increasing at an alarming rate and getting out of control. Most of the malware enters the system when we are downloading files from internet. Currently, the most important line of defense against malware is internet security software products, which mainly use a signature-based method to distinguish threats in the clients. Given a collection of malware samples, these techniques first categorize the samples into families so that samples of same family distribute some common traits and generate common strings to detect variation in family of malware samples. In recent year of malware detection many techniques are applied on malware categorization such as Support vector Machine, Decision Tree and Random forest, Ensemble clustering, OOA fast FP growth, Ensemble of Classification, Single class learning Method, Flow Graph Machine algorithm etc.

Phishing website categorization - Phishing website fraud is comparatively new internet crime to malware attack. Phishing is a form of online fraud such as social engineering schemes by sending e-mails, instant message or online advertising attracts users to phishing website that pretend to be trustworthy website in order to trick individuals to help their sensitive information for example- financial accounts, account password, and personal identification numbers. This is used for profit. Phishing is the website to retrieve sensitive information such as account password, user names and user ID, credit card details and indirectly money masquerading through electronic communication. Phishing is a repetitive threat that keeps growing these days. The danger is growing larger in social media such as Twitter, Facebook, and MySpace etc. Hackers commonly use these sites to attack persons using media sites at their place of work, homes, or in public to take personal and security information which can affect the user and the company. To defend against phishing websites, security software products use blacklisting to filter against known websites. In recent years many techniques for phishing website detection are defined such as k-means clustering, Associative Classification, Neuro Fuzzy Scheme, Association Rule-based Data Mining, Classification, Ensemble clustering and other machine learning method.

In this paper we propose Kernel k-means clustering algorithm which is categorize not only malware sample but also phishing website. Kernel k-means is an extension of the k-means algorithm where vectors are mapped from vector space to a higher dimensional feature space through kernel function and then applied k-means in feature space. Thus kernel k-means avoids separable clusters in vector space and improve the accuracy of phishing website and malware categorization. The malware and phishing website categorized on the basis of feature vector TF (term frequency) and TFIDF (term frequency inverse database frequency). In this paper we introduce two main contributions (1) Feature extraction for malware and phishing website (2) showing kernel k-means clustering is capable of detecting malware and phishing website.

The rest of the paper is organized as follows. In section 2 we give details of related work. In Section 3 we describe system architecture. Section 4 describes the feature extraction and representation of phishing websites as well as malware sample. Proposed methodology described in section 5. Section 6 describes result and analysis work. Conclusion and future work are discussed in section7.

2. RELATED WORK

Yanfang et al. [1] in this paper developed an intelligent malware detection based system using objective-oriented association mining classification which consists of three stage process 1) PE parser 2) OOA rule generator 3) rule based classifier. In OOA fast FP growth algorithm, efficiency is modified to generate OOA rules to classification. But this system only provides binary predictions, that is a PE file is malicious or not.

Menahem et al. [2] in this paper evaluate several combining processes using five different base inducer (Navive Bayes, KNN, C4.5 Decision tree, VFI and OneR) on five type malware dataset. The aim is to search best combining method for the task of detecting malicious software in terms of accuracy. They measure three parameter accuracy, execution time, and AUC. This classification Technique requires large number of training dataset to perform the classification model.

Ye Li et al. [3] in this paper Malware samples are detected on the basis of instruction frequency and function based instruction sequence feature vector. They develop an automatic malware categorization based system for automatically detected malware samples into families that share some similar properties using cluster ensemble from aggregating the clustering solution generated by different clustering. So there are many clustering algorithm perform after which connectivity matrix applied to generate aggregation solution.

Zhao et al. [4] in this paper researched on software structure and found that control flow of software divided into many basic block from interior cross-reference and the feature selection method can extract opcode sequence from dissembled program and then rules may be applied on malware detection. This malware detection method substitutes categorize rules for signature which was widespread used in traditional malware detection process which becomes too difficult to understand specific meaning of the class if rules due complexity.

Santos et al. [5] the author proposed a new technique which uses single-class learning to detect unknown malware samples. The method is based on analysis the frequencies of the appearance of opcode sequences to perform a machine learning classifier using only a set of labeled instance within a specific class malware or legal software. This method can reduce the effort of the labeling for software and maintaining high accuracy. It is difficult to obtain large amount of labeled data in real world environment.

Aburrous et al.[6] in this paper present the novel approach to overcome the difficulty and complexity in detecting and predicting phishing website. They proposed an intelligent resilient and effective model that is based on classification and association data mining algorithm. This algorithm used to classified phishing website and relationship with them.

Liu et al. [7] in this paper the approach to identification of the phishing website target of given webpage is clustering the webpage set consisting of its all associated webpages and given webpage. They first found associated webpages, and then extend their relationships to webpage for feature extraction and then applied DBSCAN clustering algorithm.

Abdelhamid et al. [8] in this paper investigate the problem of phishing website using a AC method that is called Multi label classifier based associative classification (MCAC) to search its applicability to phishing problem. MCAC generate new hidden knowledge which other method unable to found to improve its classifier predictive performance.

Wenyin et al. [9] the authors develop anti-phishing classify phishing technique which uses visual characteristic to classified potential phishing website and measure suspicious webpage similarity to real sites registered in system. First two sequential methods in the sites system run on local mail server and monitors mails for keywords and suspicious URL. Second method compare the potential phishing webpage against real pages and assess visual similarities between them in the terms of key regions, page styles and other styles.

Zhuang et al.[10] Clustering ensemble is to the process of obtaining a single or consensus and better-performing clustering solution from a number of different input clustering's for a particular dataset. Many techniques have been developed to explain ensemble clustering problems over the past few years. However, most of these methods are designed to combine partitional clustering methods, and few have been detected for combining both partitional (k-medoid) and HC (agglomerative) methods. In addition, they don't take advantage of domain-related constraints. In the study, they applied a cluster ensemble to combined the clustering solutions which are generated by both hierarchical (agglomerative) and partitional (k-medoid) clustering methods.

3. SYSTEM ARCHITECTURE

Fig-1 Show the architecture of the detection based on kernel k-means clustering and we briefly describe each component below.

Term-frequency feature extractor: In phishing website categorization, the system first uses the term-frequency feature extractor to extract the terms form the WebPages of the collected phishing website and transforming the data into term- frequency feature vector. These vectors are stored in database. Transaction data can be simply converted to relational data if necessary.

Instruction-frequency feature extractor: In malware categorization, firstly system uses the instruction- frequency feature extractor to extract the function-based instructions from collected portable executable (PE) malware samples. These integer points or vectors are transformed to instruction frequencies and stored in the database. The transaction data can also be simply converted to relational data if necessary.

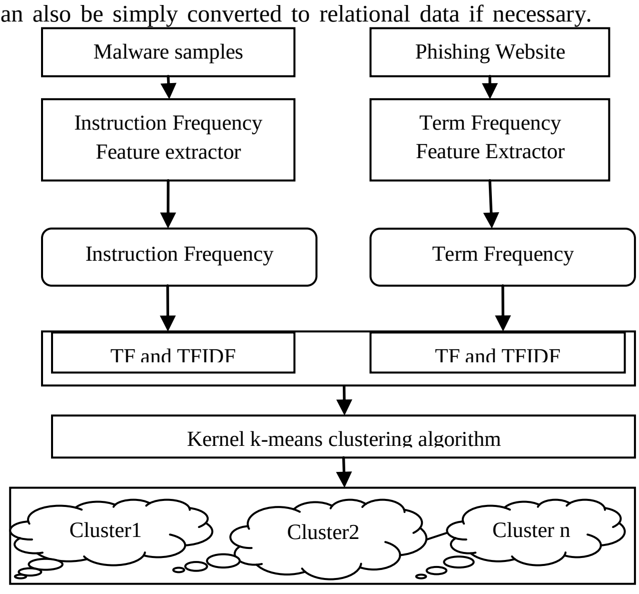

Fig 1: Proposed Architecture

Clustering algorithm: clustering solutions are generated by applying kernel function based k-means clustering algorithm which is based on feature representation. In the kernel k-means clustering vectors are mapped from vector space to a higher dimensional feature space through kernel function and then applied k-means in feature space.

4. FEATURE EXTRACTION

Han et al. [21] a set of d document and a set of t terms, which model each document as a vector v in t dimensional space R^t, this method is called integer input space model. The freq (d,t) is frequency of the number of occurrences of term t in document d. TF (term frequency) is basically number of times a given term appears in specific document. After that term frequency measure there is another important measure, it is called IDF (inverse document frequency). That is representing the scaling factor and importance of a term t, if a term t occurs in many document, its importance will be scaled down due to its reduced discriminative power. In a complete vector space model that is TF and IDF are combined together which forms the TFIDF. Thus the term has a high TFIDF weight having high term frequency in given document.

If freq(d,t) = 0 then that is TF(d,t) = 0

Otherwise TF(d,t) = 1+log(1+log(freq(d,t)))

IDF(t) = log $(1+|d|)/|d_t|$

TF-IDF(d,t) = TF(d,t) * IDF(t)

Where d_t is set of document containing term t.

4.1 Feature Extraction of Malware Samples

There are mainly two ways for feature extraction in malware analysis these are static and dynamic feature extraction present behaviors of malware files and perform well in analyzing packed malware Bailey et al. [11] and Lee et al. [12]. However, it has limited executable files which can be executed or simulated. Data collected from the Malware Research & Data Center cannot be dynamically analyzed. On the other hand the dynamic feature extraction is time consuming. Therefore, in our study, we choose static feature extraction method for malware representation. We use the instruction frequencies for malware representation for comparing with other static feature Tian et al. [13], such as construction phylogeny tree, control flow graph. Windows API calls and arbitrary binaries, the instruction frequencies or function-based instruction sequence for malware representation have great ability to represent variants of a malware family, high coverage rate of malware samples, good semantic implications, and high efficiency for feature extraction. T et al. [14] the extraction and transformation processes are given in following steps.

- Firstly select the program executable (PE) file for malware sample. Now we have list of all assembly instruction. Compared these instructions with program executable code.
- The list of instruction prepared with instruction frequency.
- Convert instruction frequency into TF and TFIDF.

4.2 Feature Extraction of Phishing Websites

There are several feature extraction methods for phishing website representation URL of the website Chou et al. [14], user interface, Wu et al. [15], associated WebPages of the website G. Liu et al. [8], webpage block, page layout, and whole style W. Liu et al.[9], term f given webpage with the TF-IDF score Dazeley et al.[16] Zhang et al. [17] etc. considering the expression ability of the website and the complexity for the categorization inputs. In this paper, we extract the term frequencies from the WebPages of their relating to website. We firstly extract the terms form the website html pages on basis of java script. The description of the extraction is illustrated as follows.

- In the first step select script between tags in the HTML file, this is in the code form.
- Transform the all script in the upper case alphabets. Then remove the extra symbols except the upper case alphabets.
- After that define the minimum word length l_1 and maximum word length l_2. And removing all the extras word, which are not between l_1 and l_2.
- Then let there be a fix term frequency f which can be greater than f in files.
- Convert these term frequency into TF and TFIDF.

5. PROPOSED METHOD

In this paper, we proposed Kernel k-means clustering to categorize malware and phishing website. Kernel k-means Tzortzis et al. [18] is a generalization and also standard k-means clustering algorithm where vector are mapped from vector space to a higher dimensional feature space through a kernel function and then k-means is applied into feature space. The outcome of the kernel k-means clustering algorithm is in linear separators of feature space which exchange in to nonlinear separators in vector space. Thus, kernel k-means avoids the disadvantage of linearly separable clusters in vector space which is not in k-means. The objective function that kernel k-means minimize clustering error in feature space. We can identify a kernel matrix $K\epsilon R^{N\times N}$, where $K_{ij} = Ø(x_i)^T Ø(x_j)$ and taking advantage of the kernel trick, we can calculate the squared Euclidean distances in (2) without explicit knowledge of the transformation using (3) any positive-semi definite matrix (PSD) can be used as a kernel matrix. Observe that in this case cluster centers m_k in feature space cannot be calculated. In the kernel k-means clustering, the kernel function $K(x_i, x_j)$ is used to directly present the inner products in feature space without explicitly defining transformation, therefore $K_{ij} = K(x_i, x_j)$. Kernel k-means is described in Algorithm.

Algorithm : Kernel *k*-Means

Input: Kernel matrix K, Number of clusters k, Initial clusters $C_1,\dots,C_k$

Output: Final clusters $C_1,\dots,C_k$ Clustering error E

For each point x_n and every cluster Ci compute.

$\|Ø(x_n)-m_i\|^2$ using (3)

2. Find $c^*(x_n) = \text{argmin}_i (\|Ø(x_n)-m_i\|^2)$

3. Update clusters as $C_i = \{x_n \mid c^*(x_n)= i\}$4. If not converged go to step 1 otherwise stop and return.

5. Final clusters $C_1,\dots,C_k$ and E calculated using (2).

6. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we perform two sets of experimental studies over data collection obtained from malware data center [19] and phishload[20] to evaluate the categorization methods that we proposed in this paper. The first set of experiments is to get term frequency feature of the phishing website and we attain proposed approach for categorization. The second part of experiments is based on the rest analysis of instruction-frequency feature extracted from the malware sample and we attain proposed approach for categorization. All the experimental studies have been done under MATLAB tool and the environment of windows 7 operating system with Intel core i3 Processor, 4GB RAM, 500 GB HDD.

6.1 Evaluation of the Proposed Method for Phishing Website Categorization

Using phishing websites and their corresponding 1500 webpage's collection obtained from the Phishload - Tables explained.html. We construct kernel k-means clustering on TF and TF-IDF. We examine that the phishing website categorization result better then comparatively to ensemble clustering algorithm. It should be pointed out that in some cases, categorizing a phishing website to a certain family is still the prerogative of Internet security experts. For example, some of the phishing websites are prize-winning fraud websites and share similar shape of term-frequency patterns, thus may be categorized to the identical family, according to their exact intents divided into different families. On the contrary, there are some metamorphic phishing websites, like selling and social networking sites fraud which may differ from term representations but they are in the same family. The result of the kernel based system compare with Ensemble clustering (Hierarchical clustering and k-medoid) ACS [10] is given below. The result shows in table 1 and table 2 of the ensemble clustering and the proposed method on basis of the accuracy and error rate for phishing website categorization. In the graphical representation of the result shows in figure 2 and figure 3. Accuracy which is correctly classified phishing website and error rate which is incorrectly classified phishing website.

Table.1 Comparisons of Previous and proposed for Phishing website categorization on the basis of Accuracy.

No. of Phishing website	Accuracy EC	Accuracy Prop	Increases (%) in accuracy
500	37.59	50.84	13.25
1000	51.32	53.62	2.3
1500	39.14	44.34	5.2
2000	42.1	43.58	1.48

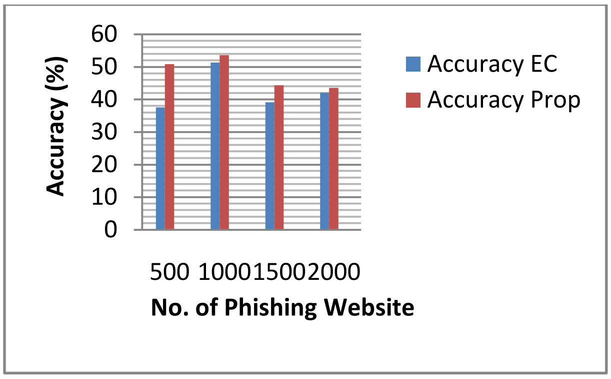

Fig.2 Comparisons to Ensemble clustering and kernel k-means clustering for phishing website categorization.

Table.2 Comparisons of Previous and proposed for Phishing website categorization on basis the of Error rate.

No. of Phishing Websites	Error rate (%) EC	Error rate (%) Prop	Decreases % in error rate
600	61.0	44.95	16.1
800	50.48	44.97	5.5
1200	56.96	56.12	0.78
1400	56.75	55.97	0.8

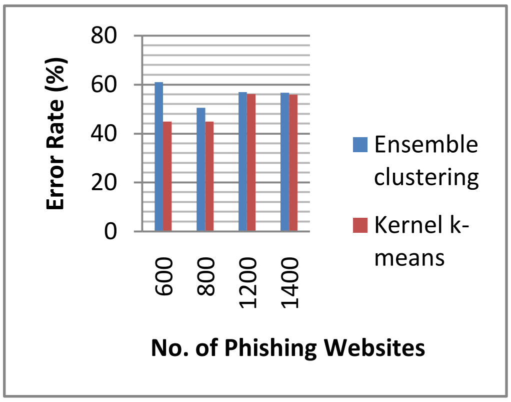

Fig.3 Comparisons to Ensemble clustering and kernel k-means clustering for phishing website categorization

6.2 Evaluation of the proposed method for Malware categorization

In this section, it is based on the daily new malware sample collection obtained from the Malware Research & Data Center. Evaluate the effectiveness of malware categorization results of our proposed clustering method compared our system for malware categorization with Ensemble clustering (Hierarchical clustering and k-medoid) ACS system [10]. Evaluation effectiveness of malware categorization results in kernel k-means clustering algorithm from ensemble clustering algorithm result shows in table 3 and table 4 of the ensemble clustering and the proposed method on basis of the accuracy and error rate for Malware sample categorization. In the graphical representation of the result shows in figure 4 and 5.

Accuracy which is correctly classified malware sample and error rate which is incorrectly classified malware sample.

Table.3 Comparisons of Ensemble clustering and Proposed for Malware sample categorization.

No. of Malware sample	Accuracy EC	Accuracy Prop	Increases (%) in accuracy
10	58	78	20
20	42.3	50.13	7.83
30	37.93	44.77	6.84
40	41.89	48.31	6.42

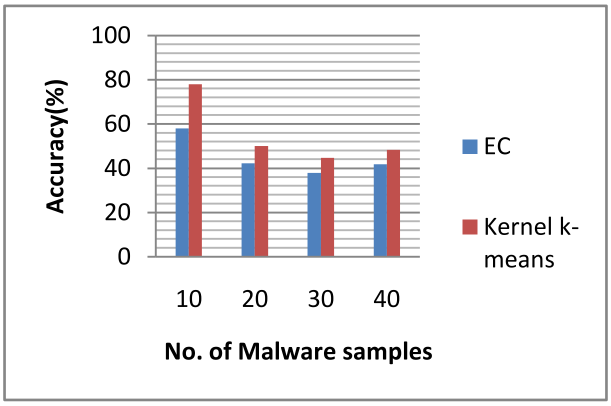

Fig.4 Comparisons of Ensemble clustering and Proposed for Malware categorization

Table.3 Comparisons of Ensemble clustering and Proposed for Malware sample categorization.

No. of Malware sample	Error rate EC	Error rate Prop.	Decreases (%) in error rate
15	46.67	28.4	18.27
25	50.00	46.89	3.11
35	46.00	36.67	9.33
45	52.00	45.00	7.0

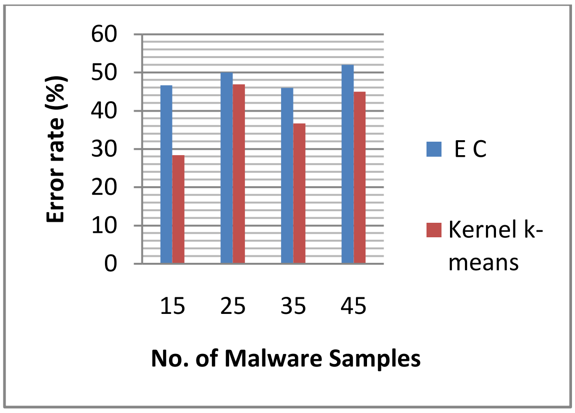

Fig.5 Comparisons of Ensemble clustering and Proposed for Malware categorization

7. CONCLUSION

In this paper, firstly extract feature from the phishing website and malware sample through term frequency and instruction frequency then we have developed a system which is applied for malware sample categorization or phishing website categorization into families that share some common traits by kernel k-means clustering method. The studies on large and standard data set collected from malware data center and Phish load - Tables explained.html show that our system performs well for real phishing website categorization as well malware categorization applications. In this technique the categorization accuracy and error rate of phishing website or malware sample improved as compare to ensemble clustering technique. The accuracy and error rate improve 10 to 20 % for both malware and phishing website categorization.

In the future work that can enhance with three fields firstly in this work the feature extracted by different method for malware and phishing website, second there are many clustering algorithm which can be applied to malware and phishing website categorization and third one is that can include anomaly detection with malware and phishing website categorization.

8. ACKNOWLEDGMENTS

I would like to thank Dr. S. K. Shrivastava, Associate Professor, Department of Information Technology, who has contributed towards development of the template.

9. REFERENCES

[1] Y. Ye, D. Wang, T. Li, D. Ye, and Q. Jiang, Jan. 2008 “An intelligent PE-malware detection system based on association mining,” J. Comput. Virol., vol. 4, pp. 323–334.

[2] E. Menahem, A. Shabtai, L. Rokach, and Y. Elovici, Feb. 2009 “Improving malware detection by applying multi-inducer ensemble,” J. Comput. Stat. Data Anal., vol. 53, no. 4, pp. 1483–1494.

[3] Y. Ye, T. Li, Y. Chen, and Q. Jiang, 2010 “Automatic malware categorization using cluster ensemble,” in Proc. 16th ACM SIGKDD Int. Conf. Knowl.Discovery Data Mining, pp. 95–104.

[4] Zongqu Zhao, Junfeng Wang, Jinrong Bai1 2014, “Malware detection method based on the control-flow construct feature of software” IET Inf. Secur. Vol. 8, Iss. 1, pp. 18–24 .

[5] I. Santos F. Brezo B. Sanz C. Laorden P.G. Bringas 2011 “Using opcode sequences in single-class learning to detect unknown malware” IET Inf. Secur,Vol. 5, Iss. 4, pp. 220–227

[6] M. Aburrous,M. A. Hossain, K. Dahal, and F. Thabtah, 2010 “Predicting phishingwebsites using classificationmining techniqueswith experimental casestudies,” in Proc. 7th Int. Conf. Inf. Technol. pp. 176–181.

[7] Neda Abdelhamid, Aladdin Ayesh , Fadi Thabta 2014 ”Phishing detection based Associative Classification data mining” 0957-4174/ Elsevier Ltd. All rights reserved.

[8] G. Liu, B. Qiu, and L. Wenyin, 2010 “Automatic detection of phishing target from phishing webpage,” in Proc. 20th Int. Conf. Pattern Recognit, pp. 4153–4156.

[9] W. Liu, X. Deng, G. Huang, and A. Y. Fu, Mar./Apr. 2006,“An antiphishing strategy based on visual similarity assessment,” in Proc. IEEE Internet Comput, pp. 58–65.

[10] Weiwei Zhuang, Yanfang Ye, Yong Chen, and Tao Li Nov 2012 ” Ensemble Clustering for Internet Security Applications” IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 42, NO. 6.

[11] M. Bailey, J. Oberheide, J Andersen, Z. M. Mao, F. Jahanian, andJ. Nazario, 2007 “Automated classification and analysis of internet malware,” in Recent Advances in Intrusion Detection, (Lecture Notes in Computer Science vol. 4637). New York: Springer, pp. 178–197.

[12] T. Lee and J. J. Mody, May 2006 “Behavioral classification,” in Proc. EICAR.

[13] R. Tian, L. M. Batten, and S. C. Versteeg, 2008 “Function length as a tool for malware classification,” in Proc. 3rd Int. Conf. Malicious Unwanted Software, pp. 69–76.

[14] N. Chou, R. Ledesma, Y. Teraguchi, D. Boneh, and J. C.Mitchell, 2004 “Clientside defense against web-based identity theft,” in Proc. 11th Annu. Network Distrib. Syst. Secur. Symp.

[15] M.Wu, 2004 “Fighting phishing at the user interface” Ph.D. dissertation, Mass. Inst. Technol., MA.

[16] R. Dazeley, J. L. Yearwood, B. H. Kang, and A. V. Kelarev, 2010 “Consensus clustering and supervised classification for profiling phishing emails in internet commerce security,” in Knowledge Management and Acquisition for Smart Systems and Service (Lecture Notes in Computer Science, vol. 6232). New York, Springer-Verlag, pp. 235–246.

[17] Y. Zhang, J. Hong, and L. Cranor, 2007 “CANTINA: A content-based approach to detecting phishing web sites,” in Proc. 16th World Wide Web Conf. pp. 639–648.

[18] Grigorios Tzortzis and Aristidis Likas 2008 “The Global Kernel k-Means Clustering Algorithm” 978-1-4244-1821 3/08/$25.00© IEEE.

[19] “VirusSign” Available: http://www.VirusSignMalware Research & Data Center, Virus Free Downloads.html” © VirusSign, Inc.

[20] “Phishload” Available: http://www.Phishload-Tablesexplained.html Copyright (c) 2012Max-Emanuel Maurer (University of Munich).

[21] Han Jiawai and Kamber Micheline 2006 Data Mining concept and technique second ed. USA by Elsevier Inc.