ISSN 2250-3153

Evolutionary Fuzzy based System for Detecting Malicious Web Pages

S.Chitra, K.S.Jayanthan and S.Preetha

Abstract- Internet became a platform for fast communication and information transactions. Web pages undergo constant dynamic transformations. Due to this reason, it has become attackers preferred pathway for installing illicit products. It also made the intruders to use World Wide Web to hack or attack a client's system. Malwares may be installed to the target system to disclose the user information. Malicious web pages are created using code attacks, finding system vulnerabilities and other methods. It contains potential threats. Number of web pages and malicious content in it, is increasing proportionally. Thus, it requires a large security concern in the Internet. Various approaches have been made to detect maliciousness. Evolutionary computation reached its top notch in recent years. Fuzzy systems are used in solving issues in various application domains. This paper proposes a novel approach using genetically evolved Fuzzy rules. Every web page contains its own features which may differ for a malicious web page. Using this factor, the system is implemented.

Index Terms- Drive by-download, Genetic Algorithm, Fuzzy system, Security.

I. INTRODUCTION

With the advent and the rising popularity of networks, Internet, intranets and distributed systems, security is becoming one of the focal points of research. Web content is undergoing a significant transformation. Early web pages contained simple, passive content, while modern web pages are increasingly active, containing embedded code such as ActiveX components, JavaScript, or Flash that executes in the user's browser.

Malicious web page contains potential threats, which is a collection of scripts, foreign contents or an exploited content, added by an intruder to a web site. While performing a web service, a user will make request for a particular web site to a web server. The request is sent using communication messages. The server responds to the request by sending the requested web site to the user. It may contain malicious content or other software that install malware in the client's system. These are powerful to disclose the user's confidential informations that may include personal and professional matters without their knowledge. The vulnerabilities of web browsers and web based services have increased numerously by out numbering the conventional security concerns in the computer society. Security concern should be made mandatory for every user while accessing a web page. In order to provide this security, the threat of maliciousness will be given more priority.

The usage of internet services have become essential in today's environment. Most user interactive applications are written in JavaScript or VBScript. These scripting Languages provides an easy access to a service but because of the vulnerabilities of web browsers users security is compromised as it is easy to inject a malignant web code in to a web page since the user system is neither patched nor updated. As a matter of fact, more and more people are concerned with malicious code that could exist in software products. Malicious codes are pieces of code that can affect the secrecy, the integrity, the data and control flow, and the functionality of a system.

Most user system use anti-virus softwares to detect DHTML-code attacks. These softwares concentrate on the detection of binary executables by using signature based encryption. Their efficiency depends on the updation frequency. Changing the content of DHTML code is easier than changing binary executable programs. The updating frequency of signatures is slower than transformation frequency of malicious DHTML codes. This makes signature-based techniques ineffective to detect variants of malicious codes. In an experiment conducted, which tested the efficiency of anti-virus software against the commonly used obfuscation, the average false negative rate was 40% to 80%.

This paper proposes a method for detecting maliciousness using genetically evolved fuzzy rules. The potential features of a web page are extracted in this paper for implementation.

II. LITERATURE REVIEW

To detect malicious web pages, the approaches used are mainly classified as: State change technique, Signature technique and Machine learning approach. The major techniques used by an attacker on a system are SQL injection, Obfuscation, Browser vulnerability Exploitation and URL redirection. The popular search engines are being abused by compelling the users to visit malicious web sites.

State change approach which is also known as rule-based approach monitors change of state against unauthorized creation of executable files or register entries in the client system[29].To detect drive by-download attack Bragin *et al* used event triggers by creating some trigger conditions to find unauthorized activities in file system, process creation and registry system[6]. Behaviour monitoring module is also conducted in a client system to track malicious behaviour [23].

To detect malicious web pages, Bin, Jianjun, Fang, Dawei, Daxiang and Zhaohui [13] proposed the concept of abnormal visibilities. The authors showed three main forms of abnormal visibility: width and height attributes of iframe, setting the display style of iframe 'display: none', generating iframe tag dynamically to make obfuscation. Abnormal visibility fingerprints are created and used to detect malicious web pages. Each web page is scanned to detect any form of abnormal

visibility. The detected value in any kind of abnormal visibility is compared with a threshold value.

SQL injection exploits a vulnerable database application that runs in web servers. It allows unauthorized operations in vulnerable databases to collect user information and also to make changes in the data itself. This allows the adversary to directly alter the contents of the server's database and inject his own content [22].

Signature system uses low or high interaction client side Honeypot. It uses signatures for finding the malicious web page. In HoneyC system, Snort signature is used and in Monkey-spider system, contents of web page are crawled and stored in files [1]. Unknown attacks are not considered in this approach which becomes its drawback.

Machine learning approach consists of various methods to find maliciousness. HTTP responses from potential malicious web pages are analyzed to extract potential malicious characteristics [3], [21]. Another method proposes finding of malicious page by choosing features according to the usage of DHTML [30]. A Semantic aware reasoning detection algorithm based on structures of HTML codes is malicious web proposed to detect page [16]. Features of web pages, differentiated as static and run time are been selected to check the maliciousness of a web page using scoring algorithm [28].

Thus most of the machines learning approaches use features that are extracted from web page properties. Compromising a client system or a legitimate web site causes spread of malicious content. Thus this is not enough to classify malicious web page from a benign page.

III. METHODOLOGY

This paper proposes a method to find malicious web page using genetically evolved fuzzy rules. Potential features of a web page are selected. These features are those that remain unchanged in a web page while it is executed. [28] Van Lam Le, Ian Welch, Xiaoying Gao, Peter Komisarczuk suggested 26 potential features from which we selected 21 optimal features.

TABLE 1
POTENTIAL FEATURES

Sl.no.	Features
1.	Number of redirection
2.	Number of iframe and frame tag
3.	Number of external link in iframe and frame tag
4.	Iframe and frame link length
5.	Ratio of vowel character in iframe and iframe link
6.	Ratio of special character in iframe and frame link
7.	Number of external links(other than iframe)
8.	Other link length
9.	Number of scripts
10.	Number of script lines
11.	Ratio of special character in script
12.	Script length
13.	Script word length
14.	Script function argument length
15.	Number of objects
16.	Number of applets
17.	Object link length
18.	Ratio of special character in object links
19.	Ratio of vowel character in object links
20.	Number of object attributes
21.	Applet link length

Binary encoding and Rule creation are the two important steps in this approach. Genetic algorithm and fuzzy rules used to make the binary encoding of features and the rule creation respectively. The binary encoded conditions undergoes cross over to develop new and more combination rules. Thus various new rules of web page features can also be checked.

Genetic Fuzzy System

It is one of the most successful approaches to hybridize fuzzy systems with learning and adaption method apart from neural network [16], [17]. Genetic fuzzy systems are soft computing paradigm which focuses on the design and generation of fuzzy rules using evolutionary algorithm. It can solve complex real world problems which are difficult to be solved by conventional systems.

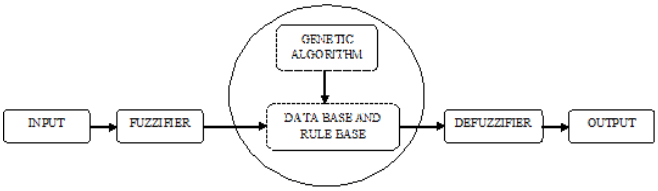

Fig 1. Genetic fuzzy System

The Michigan-style genetic fuzzy rule-based system is a machine learning system which employs linguistic rules and fuzzy sets in its representation and is ideal for the rule discovery [18]. Genetic Algorithms are search algorithms based on natural genetics that provide robust search capabilities in complex spaces and thereby offer a valid approach to problems requiring efficient and effective search processes.[1] [19].This approach mainly used in all type of probabilistic optimization problems and is inspired by biological evolution process. A Genetic Algorithm maintains a population of candidate solution for the problem at hand, and makes it evolve by iteratively applying a set of stochastic operators. The operators mainly used are mutation and crossover.

Fuzzy System

A fuzzy rule based system consists of two components: knowledge base and inference system [20]. Definition of the database and derivation of the rule base is associated with knowledge base. The definition of database is associated with variable universe, scaling factors of function, granularity per variable and membership functions associated with labels. Derivation of the rule base associated with fuzzy rule composition. The design of the knowledge base majorly focused on machine learning or human expert information.

Fuzzy logic

Fuzzy rules carries out the concept of partial truth or partial membership. There are membership values that range from 0.0 to 1.0 which denotes the false and truth respectively. Linguistic variables are essential to approximate the reasoning, because they are used to determine truth and possibility values of fuzzy propositions [21].Fuzzy logic provides an alternative way to represent linguistic and subjective attributes of the real world in computing

IV. Proposed Method

Using Genetic Algorithm, from a set of initial fuzzy rules we are creating combinations. The initial fuzzy rules used are Mamdani Fuzzy rules. As already seen, genetic algorithm is a function optimizer. It will produce the combination of fuzzy rules using mutation and cross over. Here we are using multiple-point mutation and double-point cross over. The method uses a probabilistic measure to apply genetic algorithm operators.

$$p(mutation/cross\ over) = p(mutation \cap cross\ over)/p(cross\ over)$$

For each webpage extract the basic feature vectors and calculate the antecedent occurrence frequency of the corresponding rules. That measure can be taken as the weightage of the rule and is given by

$$max\ (A/A+B, B/A+B) \quad (1)$$

A rule with 'n' features in its antecedent part, have a weightage which is given by

$$w(r) = max\left(\frac{A1}{\sum_{i=1}^{n} Ai}, \frac{A2}{\sum_{i=1}^{n} Ai}, \dots \frac{An}{\sum_{i=1}^{n} Ai}\right) \quad (2)$$

$A1, A2, \dots An$ are the feature vector frequency in the webpage.

Expectation of taking best rule,

$$r_1 p_1 + r_2 p_2 + r_3 p_3 + \dots + r_n p_n = 1 \quad (3)$$

$p(Best\ rule)$ = Rule with maximum weightage

$= w(r1)$ if $r_1 > r_2 \dots\dots r_n$

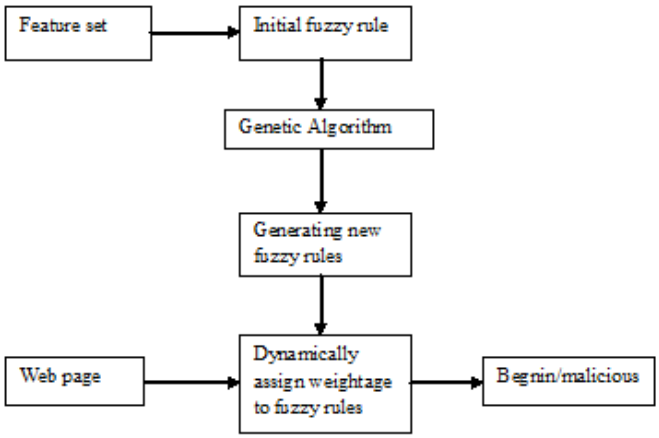

Fig 2. Architecture for proposed system

Algorithm

1. Start
2. Extract and analyze the web data set with the potential features.
3. Initialize the fuzzy rules using MAMDANI fuzzy model.
   - If Ai && Aj then Ri
   - If Bi && Bj then Rj
   - If Ci && Cj then Rk
4. Generate new featured fuzzy rules using Genetic Algorithm for a fixed number of iterations. These new rules are created by mutation and cross over,
   - A'i && B'j then R'j
   - A'j && B'i then R'i
   - A'i && C'i then R'k
   - B'i && C'j then R'i
5. Assign weightage to the initially formed fuzzy rules.
6. Apply the above formed weighted rules on the features of the webpage to confirm the maliciousness.
7. Stop.

V. Implementation

The proposed method is simulated using matlab. The potential features in Table 1 are extracted from the web pages. Using some of these features, initial fuzzy rules are developed. The genetic algorithm tool kit in matlab generates fuzzy rules. Fuzzy rules are given as the population and it will undergo mutation and cross over, to generate new combinations of rules. The rules are being represented as bit form. Each rule will be of 17 bits. Using these rules the maliciousness in the web pages are found. Considering a single web page, it may be malicious or non-malicious. In this situation, weightage is given for each potential features. Thus conformance in the malicious page detection is made.

VI. Result and Discussion

Sample formed rules:

If (number of iframe and frame tag= low) or (number of redirection= high) then malicious= high

If (number of iframe and frame tag= medium) or (number of redirection = high) then malicious = high

If (number of iframe and frame tag = low) or (number of redirection = medium) then malicious = medium

If (number of iframe and frame tag = low) or (number of redirection = low) then malicious = low

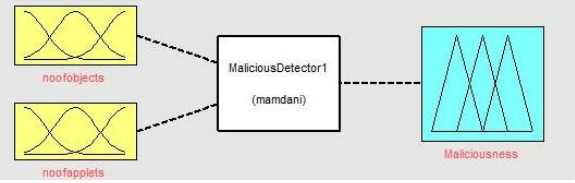

Fig 3: Fuzzy controller for Malicious Detector

The formed rules are converted into bit form in the following way. As there are 26 features, the feature vector is represented as 5 bit vector. Linguistic variables are represented using 2 bits. Output variable is represented using 1 bit. Thus the single rule will be of 17 bits that includes two features and result.

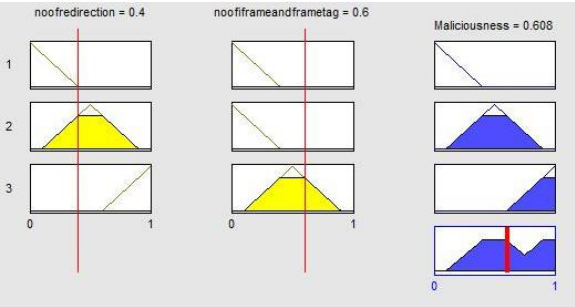

Fig 4: Fuzzy rule with membership value

For each rule we have to allot appropriate weightage using *max (A/A+B, B/A+B)*. This Max-weight technique is applied, when the page contain similar value for different features. Then the probability of selecting best rule is pointed to the rule with maximum weightage.

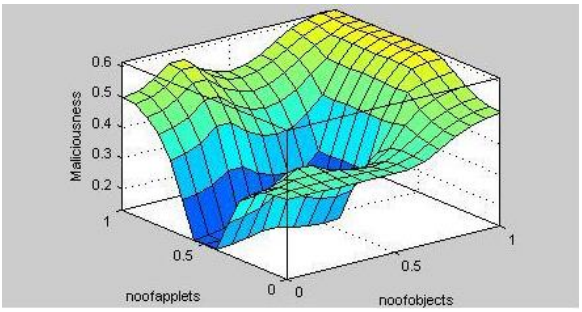

Fig 5: Fuzzy rule surface

VII. Conclusion and Future Work

The formed rules accurately measure the maliciousness of the web page. This work majorly concentrates on malicious detection using genetically evolved fuzzy rules. The weightage given to the potential features of the web page gives more conformances to the presence of maliciousness.

Inclusion of more number of potential features and type II fuzzy are further enhancement for this work. The collected malicious web pages can be classified using a non-linear classifier to increase accuracy.

References

[1] A.Ikinci, T. Holz and F. Freiling, Monkey-Spider: Detecting Malicious Websites with Low-Interaction Honeyclients, Sicherheit, Saarbruecken, 2008.

[2] C. Seifert, I. Welch and P. Komisarczuk, HoneyC - The Low- Interaction Client Honeypot, NZCSRSC, Hamilton, 2007.

[3] C. Seifert, I. Welch and P. Komisarczuk, Identification of Malicious Web Pages with Static Heuristics, Telecommunication Networks and Applications Conference, 2008. ATNAC 2008. Australasian, 2008.

[4] Chia-FengJuang, Shih-Hsuan Chin and Shu-Wew Chang, A self organizing TS-Type Fuzzy Network with Support Vector Learning and its Application to Classification problems, IEEE transactions on Fuzzy Systems, vol. 15, no.5, 2007.

[5] Davis, La Jolla, CA: Morgan Kaufmann, Adapting operator probabilities in genetic algorithms, Proceedings of the Third International Conference on Genetic Algorithms, 60-69,1989

[6] E. Moshchuk, T. Bragin, S. D. Gribble and H. M. Levy, A crawlerbased study of spyware on the Web, (2006).

[7] Francisco Herrera, Genetic Fuzzy systems: A state of Art and new trends.

[8] Genetic Fuzzy Systems: Status, Critical Considerations and Future Directions by Francisco Herrera on International Journal of Computational Intelligence Research. ISSN 0973-1873 Vol.1, No.1 (2005), pp. 59-67.

[9] Introduction to Fuzzy Systems, Neural Network and Genetic Algorithms by Hideyuki TAKAGI in Intelligent Systems: Fuzzy Logic, Neural Network and Genetic Algorithms Ch.1 pp.1-33 by D.Ruan, Kluwer Academic Publishers, September 1997.

[10] J. Ma, L. K. Saul, S. Savage and G. M. Voelker, Beyond blacklists: learning to detect malicious web sites from suspicious URLs, Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, Paris, France, 2009.

[11] Jorge Casillas, Brian Carse and Larry Bull, Fuzzy-XCS: A Michigan Genetic Fuzzy System, IEEE Transactions on Fuzzy Systems, vol. 15, no. 4, 2007.

[12] Kowaiczyk, R. On numerical and linguistic quanti_cation in linguistic approximation, IEEE International Conference on Systems, Man, and Cybernetics, 326{331.}, 1999

[13] L. Bin, H. Jianjun, L. Fang, W. Dawei, D. Daxiang and L. Zhaohui, Malicious Web Pages Detection Based on Abnormaingl Visibility Recognition, E-Business and Information System Security, 2009. EBISS '09. International Conference on, 2009, pp. 1-5.

[14] L. Shih-Fen, H. Yung-Tsung, C. Chia-Mei, J. Bingchiang and L. Chi- Sung, Malicious Webpage Detection by Semantics-Aware Reasoning, Intelligent Systems Design and Applications, 2008. ISDA '08. Eighth International Conference on, 2008.

[15] L.A. Zadeh, Fuzzy Sets. Information and Control 8, 338{353}, 1965.

[16] N. Provos, P. Mavrommatis, M. Abu and R. F. Monrose, All your iframes point to us, Google Inc, 2008.

[17] O.Cordon, F. Gomide, F.Herrerea, F.Hoffman and L. Magdalena "Ten years of genetic fuzzy systems: Current Framework and new trends", Fuzzy Sets Syst., vol. 141, no. 1, pp. 5-31, 2004

[18] O.Cordon, F.Herrerea, F.Hoffman and L. Magdalena, Genetic Fuzzy Systems. Evolutionary tuning and learning of Fuzzy Knowledge bases, ser. Advances in Fuzzy Systems – Applications and Theory Series. Singapore: World Scientific, 2001, vol. 19

[19] On Advantages of Scheduling using Genetic Fuzzy Systems by Carsten Franke, Joachim Lepping and Uwe Schwiegelshohn

[20] Ossi Nykanen, An Approach to Logic Programming with Type-1 Fuzzy Models Using Prolog, IADIS International Conference Applied Computing, 2006.

[21] P. Liu and X. Wang, Identification of Malicious Web Pages by Inductive Learning, Proceedings of the International Conference on Web Information Systems and Mining, Springer-Verlag, Shanghai, China, 2009.

[22] P. Niels, R. Moheeb Abu and M. Panayiotis, Cybercrime 2.0: When the Cloud Turns Dark, Queue, 7 (2009), pp. 46-47.

[23] S. Xiaoyan, W. Yang, R. Jie, Z. Yuefei and L. Shengli, Collecting Internet Malware Based on Client-side Honeypot, Young Computer Scientists, 2008. ICYCS 2008. The 9th International Conference for, 2008, pp. 1493-1498.

[24] Spears, W. M. & Anand, V., Charlotte, NC: Springer-Verlag, A study of crossover operators in genetic programming. Proceedings of the Sixth International Symposium on Methodologies for Intelligent Systems, 409-418, 1991.

[25] Syswerda, G., Vail, CO: Morgan Kaufmann, Simulated crossover in genetic algorithms. Proceedings of the Foundations of Genetic Algorithms Workshop, 1992

[26] Technical Report on Ten Lecturers on Genetic Fuzzy Systems by Ulrich Bodenhofer, Francisco Herrera. Revised version of lecturer notes from "Preprints of the International Summer School: Advanced Control-Fuzzy, Neural, Genetic", R.Mesiar, Ed.Slovak technical University Bratislava 1997. Pp. 1-69, ISBN.

[27] Time Complexity Ananlysis of Genetic – Fuzzy System for disease diagnosis by Ephzibah E.P. in Advanced Computing: An International Journal) ACIJ), Vol.2, No.4, July 2011.

[28] Van Lam Le, Ian Welch, Xiaoying Gao, Peter Komisarczuk, Two-stage classification model to detect malicious web page, International Conference on Advanced Information Networking and Applications,2011.

[29] Y.-M. Wang, D. Beck, X. Jiang and R. Roussev, Automated Web Patrol with Strider HoneyMonkeys: Finding Web Sites that Exploit Browser Vulnerabilities, IN NDSS (2006).

[30] Y.-T. Hou, Y. Chang, T. Chen, C.-S. Laih and C.-M. Chen, Malicious web content detection by machine learning, Expert Systems with Applications, In Press, Corrected Proof (2009).

AUTHORS

First Author – S.Chitra, chitraskariah@gmail.com
Second Author – K.S.Jayanthan, jayanthhere@gmail.com
Third Author – S.Preetha, preetha.ss@gmail.com