International Journal of Computer Architecture and Mobility
(ISSN 2319-9229) Volume 1-Issue 4, February 2013

An Ensemble Approach for Cyber Attack Detection using SVM

Chandrapal Singh Dangi[#1], Prof. Ravindra Kumar [#2], Prof. Gajendra Singh Chandel [#3]

[#1]M.Tech (CSE) Scholar, SSSIST Sehore, M.P.
[#2]Assistant Professor, Deptt. of CSE, SSSIST Sehore, M.P.
[#3]Head of CSE, SSSIST Sehore, M.P.
{[#1]chandrapaldangi@gmail.com, [#2]ravindra_p84@rediffmail.com, [#3]gajendrasingh86@gmail.com}

ABSTRACT:

Network and system security is of paramount importance in the present data communication environment. Hackers and intruders can create many successful attempts to cause the crash of the networks and web services by unauthorized intrusion. New threats and associated solutions to prevent these threats are emerging together with the secured system evolution. Cyber Attack Detection Systems (CADs) are one of these solutions. The main function of Cyber Attack Detection System is to protect the resources from threats. It analyzes and predicts the behaviors of users, and then these behaviors will be considered an attack or a normal behavior.
Here in the proposed work a technique of detecting malicious Socket address (IP Address and port no.) And suspicious URLs has been presented, which detects and blocks if any suspicious cases are found and passes the contents to concern user. Here we use SVM technique for classification, detection and prediction of Blacklisted IP addresses and suspicious URLs. Our system can also work as a Host Intrusion Detection System (HIDS) or Network Intrusion Detection System (NIDS). The proposed algorithm provides accuracy of 96.99% and which is the best among the present systems. It is light weight system and easy to implement on existing applications.

Keywords

Blacklisted IP, Blacklisted Port, Blacklisted Socket, Malicious URL, HIDS, NIDS, SVM

1. INTRODUCTION

Internet looks like a web of unknown routes or path, numbers of methods or ways are available for accessing web application, its contents, internet application follows OSI model, where each layer has various Protocols, security mechanism, filtering, and encapsulation. But various route means various point for malicious injection by attacker to create injury in web contents. Cyber attacks [1] are actions that attempt to bypass security mechanisms of computer system. A number of cyber attack detection and classification methods have been introduced with different levels of success that is used as a countermeasure to preserve data integrity and system availability from attacks. Cyber Security [2] concepts are studied for protecting web application from malicious data injection or from the occurrence of fraudulent scheme. Designs, tools, Algorithm and architectural testing etc. Here security for web application contents is presented by using Machine learning techniques.

Frameworks are being designed for the purpose of by detecting, checking and blocking attack signatures and attack procedure and patterns. Attacker's work by noticing or by finding bypassing mechanism to access secure connections and designs. Every attack have their predefined levels, steps and pattern but they grows as the security implication increases, as security increases ,attack potential also increases. Every web-browser has their own deign Pattern and algorithms like internet explorer, Opera, Google chrome, Netscape navigator, Mozilla Firefox, some of them works on HTTP and some on HTTPS. Https provides secure channel (Contains encryption, key exchange or algorithms schemes for packets) for web application and are less susceptible to attack but http (Packet is transferred from source to destination in original form without any encryption or key exchange) is susceptible to attack.
Example:-www.Student.com/data.aspx/id/454/branch
198.23.23.43 6777
(Original URL (IP) and Port no.)

User's are free to enter anything or supply any combination of letter's , string's or number's ,the above example contains original URL and Original IP, The below example shows edited data on the URL's contents ,which will lead to unwanted behavior in applications .The malicious IP try to create a fraud connection for the purpose of misguiding or generating suspicious packets.
Ex: www.student.com/@#$%^&*+?:{}|<>/id
198.23.23.42 3254
(Malicious URL (IP) and Port No.)

1.1 Web application Interaction is as follows

a) Web application is requested through a web browser by a user.
b) The HTTP or HTTPs protocol accepts a request of user and sent to the targeted web server.
c) Request received is executed by Server.
d) Output is generated by Application program and sent back to the user via HTTP or HTTPs.
e) Cookies maintain Current states of User, Web server and their execution report.

1.2 Support Vector Machine

The term SVM [3] is typically used to describe classification with support vector methods and support vector regression is used to describe regression with support vector methods. SVM (Support Vector Machine) is a useful technique for data classification.
Support Vector Machines (SVM) is an innovative approach to constructing learning machines that minimize generalization error. The classification problem can be restricted to consideration of the two-class problem without loss of generality. In this problem the goal is to separate the two classes by a function which is induced from available examples. The goal is to produce a classifier that will work well on unseen examples, i.e. it generalizes well. Consider the example in figure 1. Here there are many possible linear classifiers that can separate the data, but there is only one that maximizes the margin (maximizes the distance between it and the

nearest data point of each class). These linear classifiers termed the optimal separating hyper plane. Intuitively, we would expect this Boundary to generalize well as opposed to the other possible boundaries. There are several types of support vector models including linear, polynomial, RBF and sigmoid[23]

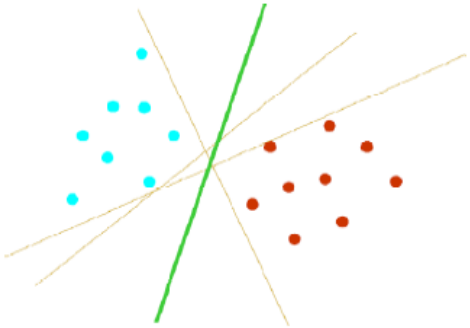

Fig. 1 Optimal Separating Hyper Plane

A classification task usually involves with training and testing data which consist of some data instances. Each instance in the training set contains one ―target value" (class labels) and several ―attributes" (features). The goal of SVM is to produce a model which predicts target value of data instances in the testing set which are given only the attributes.
To attain this goal there are four different kernel functions.

1. Linear: $K(x_i, x_j) = x_i^T x_j$.
2. Polynomial: The polynomial kernel of degree d is of the form.
$$K(x_i, x_j) = (x_i, x_j)^d$$
3. RBF: The Gaussian kernel, known also as the radial basis function, is of the form
$$K(x_i, x_j) = \exp(-\frac{||(x_i, x_j)||}{2\sigma^2})$$, where σ stands for the window width.
4. Sigmoid: The sigmoid kernel is of the form
$$K(x_i, x_j) = tanh(K(x_i, x_j) + r)$$

RBF is a reasonable first choice. The RBF kernel nonlinearly maps samples into a higher dimensional space, so it, unlike the linear kernel, can handle the case when the relation between class labels and attributes is nonlinear. Furthermore, the linear kernel is a special case of RBF show that the linear kernel with a penalty parameter C has the same performance as the RBF kernel with some parameters (C, r). In addition, the sigmoid kernel behaves like RBF for certain parameters.
. Kernel representations offer an alternative way of projecting the data into a high dimensional feature space to increase the computational power of the linear support vector machines. SVMs can classify data separated by non-linear boundaries.

2. RELATED WORK

Researchers such as Zanero [4], Kayacik [5] and Lei [6] find the Artificial Neural Network (ANN) approach more appealing. These researchers had to overcome the curse of dimensionality for the complex systems problem Liberios [7] The main goal of using the ANN approach is to provide an unsupervised classification method to overcome the curse of dimensionality for a large number of input features. Since the system is complex and input features are numerous, clustering the events can be a very time consuming task. In the computer networks cyber attack detection problem area, the size of the feature space is obviously very large. Once the dimensions of the feature space are multiplied by the number of samples in the feature space, the result will surely present a very large number. This is why some researchers, Srilatha Chebrolu[8], Gopi K.Kuchimanchi[9] and S. Selvan [10] either select a small sampling time window or reduce the dimensionality of the feature space. Since the processing time is an important factor in the timely detection of
the cyber attack, the efficiency of the deployed algorithms is very important. Time constraint may sometimes force us to have the less important features pruned (dimensionality reduction). However, the pruning approach is not always possible. Implementing data mining methodology, some researchers have proposed new data reduction approaches.

Support Vector Machine is a powerful tool to classify cyber attacks. But still it has some drawback. The first drawback is that SVM is very sensitive for attacks [11].The second, SVM designed for the two class problems it has to be extended for multiclass problem by choosing suitable kernel function. The performance of the SVM depends upon the kernel function. Some methods to improve the performance of SVM were proposed. Fuzzy SVM [12] is one of the improvements made on the traditional SVM. Several machine learning paradigms including Artificial Neural Network [13], Linear Genetic Programming (LGP) [14], Data Mining [15], etc. have been investigated for the classification of cyber attack. Also the machine learning techniques are sensitive to the noise in the training samples. The presence of mislabeled data if any can result in highly nonlinear decision surface and over fitting of the training set. This leads to poor generalization ability and classification accuracy. Decision-tree-based support vector machine which combines support vector machines and decision tree can be an effective way for solving multi-class problems. This method can decrease the training and testing time, increasing the efficiency of the system [16]. Improved Support Vector Machine (iSVM) algorithm for classification of cyber attack dataset which gives 100% detection accuracy for Normal and Denial of Service (DOS) classes and comparable to false alarm rate, training, and testing times [17][18].

A study of drive-by exploit URLs had been performed by Provos and they use a patented machine learning algorithm as a pre-filter for VM-based analysis [19]. They extract content-based features from the page, including whether I Frames are out of place, the presence of obfuscated JavaScript, and whether I Frames point to known exploit sites.

3. PROPOSE TECHNIQUE

Secure framework contains specialized feature and method's for detecting Malicious URL's and Malicious IP's .But attacker uses advance techniques to bypass their connection or query to gain access of secure information stored in server. Here a dataset Dshield and a standardized dataset is created which contains list of malicious IP address and port address with malicious query strings. Here, proposed system detects malicious URL (IP), port address and blocks them. For securing application and preventing bypassing tricks.
Ex: - Original URL (IP) and port no = www.Student.com/data.aspx/id/454/branch
198.23.23.43 1234
Malicious URL (IP) and port no = www.student.com/@#$%^&*+?:{}|<>/id
198.23.23.42 7654

Here, Malicious URL (IP) and port no. addresses are detected and blocked and Original URL and IP address are passed for processing.

Standard Dataset contains labels for detecting and blocking malicious URL (IP address) and Port no.

O-Original URL (IP address) and Port No.
Ex:- ('O')www.Student.com/data.aspx/id/454/branch 198.23.23.43 7658

M-Malicious URL (IP address) and Port No.
Ex:- ('M')www.student.com/@#$%^&*+?:{}|<>/id 198.23.23.42 9456

The above showed URL and IP address contains different combinations for accessing application by an authorized manner. Here only Original URL and IP address are passed by checking and detecting safe combinations. . Proposed system can also work as a Host Intrusion Detection System (HIDS) or Network Intrusion Detection System (NIDS) as per the users need..

3.1 Propose Algorithm

Step 1. Enter Socket address or URL in the textbox.
Step 2. Train the system from provided datasets of Suspicious Socket address or suspicious URLs(using SVM).
(i) Socket address or URLs marks as O, shows Original IP and Port address.
(ii) Socket address or URLs marks as M, shows Malicious IP and Port address.
Step 3.Predict the attack.
(i) Classify the attack using labels O (Original)
Step 4. Calculate performance and efficiency of system using labels (O and M).
Step 5.Repeat steps 1 to 3 till the correct classification Precision is achieved.
.

4. TESTING AND RESULTS

The proposed and implemented system has been tested on the datasets of Socket address (Dshield dataset [20][21]) and suspicious URLs dataset which are created. The dataset has been populated with the records of original IP Address and URLs with malicious IP and URLs was tested, whose results are calculated by taking average of 30 original data and 30 suspicious data. The system is implemented on MATLAB 7.7 for training and Classification, SVM is used for classification of original and malicious data.

TABLE -1

Original IP address and Port address (Detection Time)	Malicious IP address and Port address (Detection Time)	Accuracy
0.01432	0.01542	96.6%

Accuracy and detection time of proposed system is best among all the present system, it is light weight and easy to implement ,dataset could easily be updated by new attack patterns and ,system training could also be provided for extracting the best possible results.

Dataset of different size has been taken and accuracy is measured result is shown below in fig 3

Comparison of Accuracy and Datasize

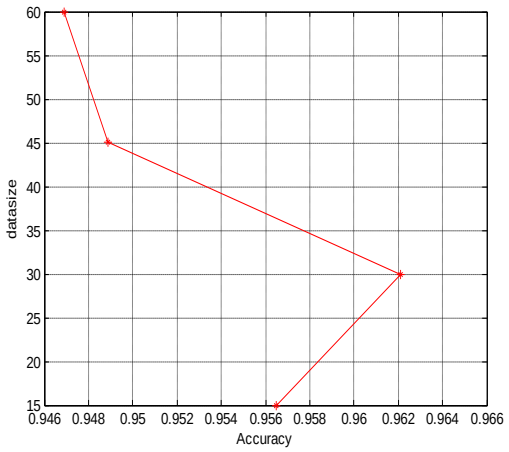

Fig 3

Different size of dataset is trained and their detection time is calculated, result is shown below in fig 4

Comparison of Detection time and Training time

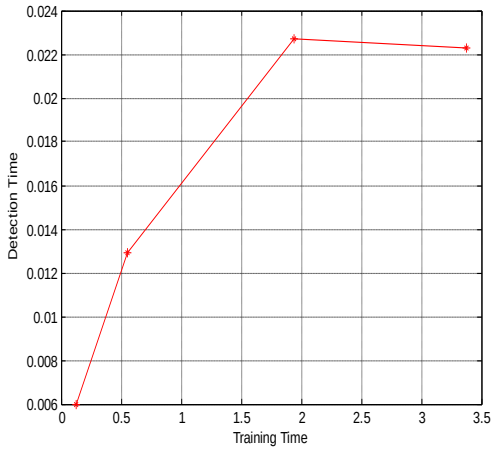

Fig 4

Accuracy and detection time is calculated when different sized dataset is used, result is shown below.fig 5

Comparison of Detection time and accuracy

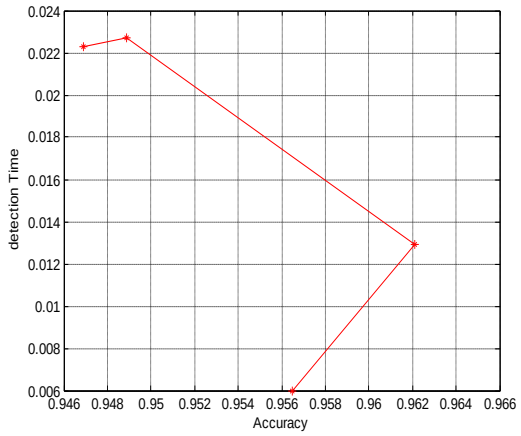

Fig 5

5. CONCLUSION

The IDS is designed to provide the basic detection techniques so as to secure the systems present in the networks that are directly or indirectly connected to the Internet. But finally at the end of the day it is up to the Network Administrator to make sure that his network is out of danger. This does not completely shield network from Intruders, but IDS helps the Network Administrator to track down bad guys on the Internet whose very purpose is to bring your network to a breach point and make it vulnerable to attacks.

We have presented an efficient technique for performing SVM especially for the case of large scale dataset where the number of training samples is large.SVM gives better detection rate, less false positive, reduced training and reduced testing times than other classifier. Here Dhsield dataset applied in the research paper used in current cyber attack detection system. Our system shows the best performance result in accuracy which is 96.6% and best among the existing systems. This can be extended by incorporating Intelligence into it in order to gain knowledge by itself by analyzing the growing traffic and learning new Intrusion patterns.

REFERENCES

[1] "A Survey of Cyber Attack Detection Systems" Shailendra Singh and Sanjay Silakari-IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.5, May 2009

[2] The Austin Forum on Science, Technology & Society Cybersecurity Today:Trends, Risk Mitigation & Research" http://www.austinforum.org/presentations/cybersecurity.pdf"

[3] Vipin Das 1, Vijaya Pathak 2, Sattvik Sharma 3,Sreevathsan 4,MVVNS.Srikanth 5,Gireesh Kumar T,‖ Network Intrusion Detection System Based On Machine Learning Algorithms‖, IJCSIT, Vol 2, No 6, December 2010.

[4] Ste. Zanero and Sergio M. Savaresi, "Unsupervised learning techniques for an intrusion detection system," in Proceedings of the 2004 ACM symposiumon Applied computing, pp. 412–419, Nicosia, Cyprus, Mar. 2004. ACM Press.

[5] H. Gunes Kayacik, A. Nur Zincir-Heywood, and Malcolm I. Heywood, "On the capability of an som based intrusion detection system," in Proceedings of the International Joint Conference on Neural Networks, vol. 3, pp. 1808–1813. IEEE.

[6]] J. Z. Lei and Ali Ghorbani, "Network intrusion detection using an improved competitive learning neural network," in Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR04), pp. 190–197. IEEE-Computer Society, IEEE, May 2004.

[7] Liberios VOKOROKOS et.al, "Intrusion detection system using self organizing map", Acta Electrotechnica et Informatica , Vol. 6 No.1, pp.1-6, 2006.

[8] Srilatha Chbrolu, Ajit Abraham, Johnson P. Thomas " Featrue deduction and ensemble design of intrusion detection systems" Computer Security, Elsevier 2004.

[9] Gopi K. Kuchimanchi, Vir V. Phoha, Kiran S. Balagani, Shekhar R. Gaddam "Dimension Reduction Using Feature Extraction Methods for Real-time Misuse Detection Systems" Proceedings of the workshop on Information Assurance and Security, US Military Academy, West Point, NY.

[10] S. Selvan, V. Venkatachalam "Performance comparison of intrusion detection system classifiers using various feature reduction techniques" International Journal of Simulation vol. 9 no.1. 2007.

[11] Liu Yi-hung, Chen Yen-ting, face recognition using total margin based adaptive fuzzy support vector machines.IEEE Transactions on Neural Networks, 18(1): 178-192.

[12] Xiong, Sheng-Wu, Liu Hong-bing, Niu Xiao-xiao, Fuzzy support vector machines based on FCM clustering. Proceddings of the fourth international conference on

[13] Machine Learning and Cybernetics, Guangzhou, China, Aug 18-21, IEEE, p.2608-2613, 2005.

[14] A. K. Ghosh and A. Schwartzbard. "A study in Using Neural Networks for Anomaly and Misuse detection" Proceeding of the 8th USENIX Security Symposium, pp. 23-36. Washington, D.C. US.

[15] Mukkamala S., Sung AH, Abraham A. Modeling Inrusion Detection Systems Using linear genetic programming approach, The 17th international conference on industrial & engineering applications of artificial intelligence and expert systems, innovation in applied artificial intelligence.

[16] W.Lee, S.J.Stolfo and K. Mok. Data mining in work flow environments: Experience in intrusion detection, Proceedings of the Conference on Knowledge Discovery and Data Mining (KDD-99), 1999.

[17] Snehal A. Mulay, P.R. Devale, G.v. Garje," Intrusion Detection System using Support Vector Machine and Decision Tree", International Journal of Computer Applications (0975 - 8887) Volume 3 - No.3, June 2010

[18] Shailendra Singh Member, IEEE, IAENG, Sanjay Agrawal, Murtaza,A. Rizvi and Ramjeevan Singh Thakur, "Improved Support Vector Machine for Cyber Attack Detection", Proceedings of The World Congress on Engineering and Computer Science 2011 Vol I WCECS 2011, October 19-21, 2011, San Francisco, USA

[19] Hoa Dinh Nguyen, Qi Cheng," An Efficient Feature Selection Method For Distributed Cyber Attack Detection and Classification",978·1-4244-9848·2/11 $26.00©2011 IEEE

[20] Justin Ma, Lawrence K. Saul, Stefan Savage, Geoffrey M. Voelker ,‖ Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs‖, *KDD'09,* June 28–July 1, 2009, Paris, France.

[21] Technical Report: Predicting future attacks http://www.ece.uci.edu/~athina/PAPERS/dshield-analysis-tr.pdf

[22] Dshield data set"http://www.dshield.org/feeds_doc.html"

[23] David Meyer, Friedrich Leisch, and Kurt Hornik. The support vector machine under test. Neurocomputing 55(1–2): 169–186, 2003