Data Preprocessing for Reducing False Positive Rate in Intrusion Detection

Dharmendra G. Bhatti
Associate Professor,
Shrimad Rajchandra Institute of Management and Computer Application,
Bardoli, Gujarat, India.

P. V. Virparia, PhD.
Professor,
Department of Computer Science and Technology,
Sardar Patel University,
Vallabh Vidyanagar, Gujarat, India.

ABSTRACT

Intrusion detection plays vital role in computer network security since long. Experience has shown that most IDS struggle for curbing false positive rate. As part of our proposed model with the objective of reducing false positive rate here we have focused on preprocessing functionality. The main objective of our preprocessing module is to reduce ambiguity and provide accurate information to detection engine. So here we have presented preprocessing module which cleans network data and handles missing or incomplete data. Preprocessing module is highly configurable. Based on the result of vulnerability assessment and network topology, hosts exists, services running, intrusion detection analyst need to configure preprocessing module. Effectiveness of preprocessing module depends on such configuration parameters and in turn knowledge of intrusion detection analyst. Preliminary analysis of our approach has shown reduction in false positive rate.

General Terms

Network Security, Intrusion Detection.

Keywords

Preprocessing, Intrusion Detection, false positive reduction.

1. INTRODUCTION

Day by day Internet and web applications penetrate through business and our lives. In today’s highly competitive market everyone wants to use Internet for their benefits. Corporate uses Internet for increasing the business by collaboration and communication. A person uses Internet for social and personal objectives. Along with great benefits Internet brings critical security issues to our doorstep. Internet becomes part of our business network in today’s communication era. Most modern businesses cannot survive or at least cannot progress effectively without Internet. On the other side attacks from Internet on business network can nullify the great benefits of Internet. If your web application is compromised, database is stolen, or server/network connectivity is subverted, the underlying business suffers critically. Intrusion detection is one of the most wanted things to run your business smoothly unless Internet is not part of your business. In collaboration with other security measures intrusion detection will provide highly needed safeguard for the business. Host based and Network Intrusion Detection Systems exists since long. Host based Intrusion Detection System (HIDS) monitors operating system level details and identifies attack on the particular host while Network Intrusion Detection System (NIDS) analyze network traffic and detect intrusion. For detecting intrusion two major approaches are in practice: signature based and behavior based. Researchers like Deris [12][25][23] have discussed hybrid Intrusion Detection Systems also. Signature based Intrusion Detection System can detect attacks with high accuracy for known attack signatures only. If new attacks are constructed from existing one, having different signature, Intrusion Detection System cannot detect that. Signature based Intrusion Detection Systems cannot detect new attacks till signature is generated and propagated to all related Intrusion Detection System implementations. Behavior based Intrusion Detection Systems are capable of identifying new attacks and many such systems are developed [4][5][25]. These Intrusion Detection Systems are trained and can identify normal behavior. If current behavior is significantly deviated from training behavior it generates an alert. It is also called anomaly based Intrusion Detection System. Anomaly based Intrusion Detection Systems are less accurate compare to signature based approach because of the overlapping normal and attack behavior. Many researchers have developed data mining based Intrusion Detection Systems [8][13][22][29][26][34][36] while others [7][17][18][33] [35][38] have used soft computing to detect intrusion. Anomaly based Intrusion Detection Systems suffer from problem of undetected intrusion and detecting normal behavior as intrusion. If one tries to increase the detection rate, it increases false positive rate respectively.

False positive alert is indentifying normal traffic as intrusion and generating alert by Intrusion Detection System. Security person overwhelmed with large number of false positives can create major problems and subsequently Intrusion Detection System will lose its credibility. Most organizations have limited human resources for network security. If such limited resources get large number of false positives, sooner or later security person will be unable to process alerts and start ignoring it. Researchers have observed that significant false positive rate not only wastes valuable resources but makes Intrusion Detection System unusable. Researchers have tried to increase the detection rate while maintaining false positive rate [1][3][36]. Just achieving high detection rate with curbed false positive rate is not sufficient. Another major challenge in intrusion detection is detect the attack online. Researchers have also worked upon how we can improve performance of Intrusion Detection System [21][31]. The root cause of problems like false negative, false positive, slow performance is raw data inputted to Intrusion Detection System. Noise, missing and incomplete data inputted to Intrusion Detection System results in false negative and false positive [1][3][36]. Duplicate and irrelevant input data increase Intrusion Detection System load and makes online detection difficult. Such problems can be restricted to a limit if not removed, by preprocessing input data [17][15][19][28][30][32][38].

2. RELATED WORK

Intrusion Detection Systems take raw network data or audit records as input, process it and identify it as normal or attack. Researchers have identified that preprocessing is needed for better results and used various approaches.

Current Intrusion Detection Systems can overwhelm with amount of information they ought to analyze. This problem is considered by Fernando [16]. They have discussed that we need to eliminate spurious and redundant information from raw data before using it for intrusion detection. They have user Rough Set for key attribute identification. Using n-gram theory they have identified redundant subsequences. They have also proposed Hidden Markov Model for service selection. Using experimental results they show how their approach reduces audit rate significantly.

A new collaborating filtering technique for preprocessing the probe type of attacks is proposed by G. Sunil Kumar [17]. They implemented a hybrid classifiers based on binary particle swarm optimization and random forests algorithm for the classification of probe attacks in a network. They used global search capability of particle swarm optimization while random forests as a classifier. Their experimental result demonstrated that as number of trees used in forest increases, the false positive rate decreases.

Frahan et al [15] has argued that all network activity is not relevant to intrusion detection. If irrelevant data is directly provided to Intrusion Detection System, it will make the detection difficult. They have discussed that selected number of features represent ad hoc network activity in better way. They have proposed feature selection algorithm in preprocessing module and claimed higher performance if irrelevant features are removed. To deal with dynamic environment of MANET they have proposed distributed and cooperative model. Through experimental results they have demonstrated effective anomaly detection with low false positive rate in ad hoc environment.

Gopi et al [19] has experimented with dimension reduction using feature extraction in misuse detection. They have used Gain in Information measure for quantitative evaluation of gain or loss in information. Two neural network methods used for feature extraction: NNPCA and NLCA. Finally they demonstrated significant test data reduction using their approach.

Preprocessing of web server log file is proposed by Shaimaa [32]. They have proposed log file preprocessing for better quality of data and consequently better mining result. They have combined different web log files with different formats in one unified format using XML. It will help in tracking extracting more attacks. Shaimaa has discussed about noisy and ambitious data of web log file and suggested preprocessing to remove such impurities. Priyanka et al [28] has also considered almost similar approach and proposed web log preprocessing for quality results.

Salem et al [30] suggested preprocessing rough network traffic in to connection records. Their tool can provide summarized and relevant information for intrusion detection. Zheng [38] has suggested Hierarchical Intrusion Detection. It uses statistical preprocessing and neural network classification. They have tested different types of neural network classifiers and also performed stress test. Sanjay [31] proposed Singular Value Decomposition as a preprocessing step to reduce the dimensionality of data. Such reduction will give importance to more prominent features in data.

3. OUR APPROACH

We have proposed a conceptual model for reducing false positive rate in intrusion detection in earlier paper [14]. Now, in this paper, we have focused on preprocessing module of our conceptual model. Objective of this preprocessing module is to reduce false positive rate.

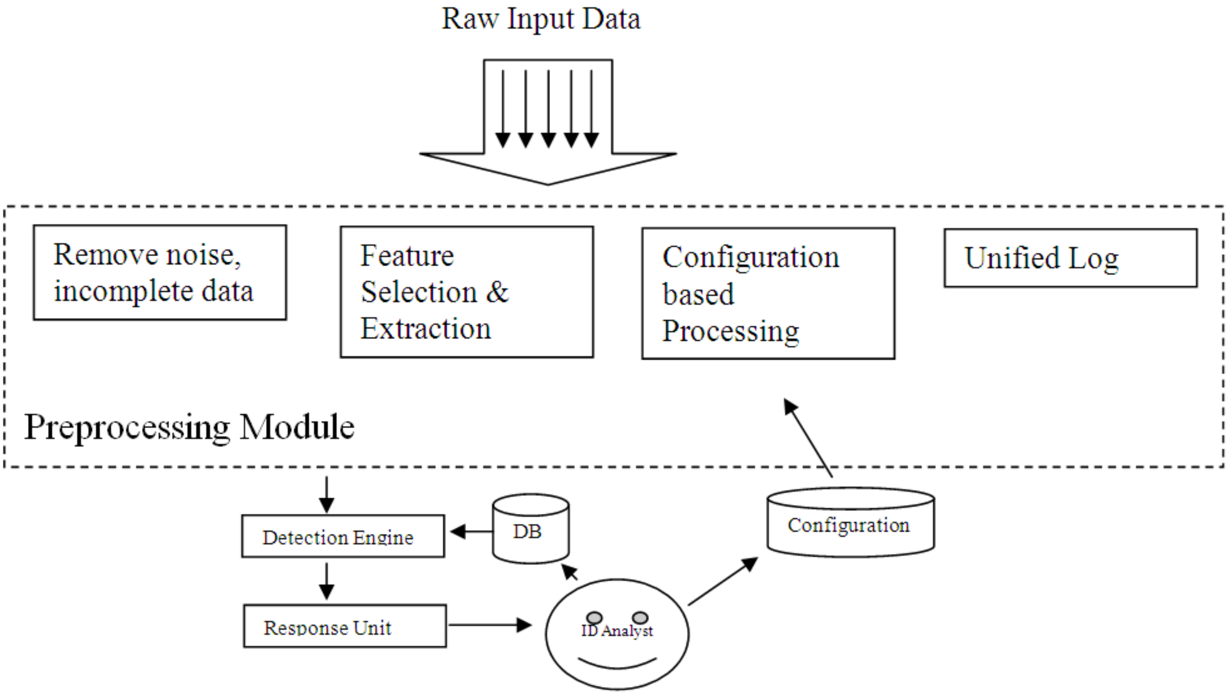

Figure 1: Data Preprocessing

We have addressed major causes of false positive as follows:

- noise, incomplete data
- spurious and duplicate data
- missing network knowledge
- multiple log formats

In the proposed model we have used four major functionalities in preprocessing module. In the first round, input data cleaning by removing noise and incomplete data is proposed. Unwanted parameters like noise and incomplete data makes the task of intrusion detection difficult. It increases overlapping behavior of normal and intrusion data. Most modern data mining [8][13][22][29][26][34][36] and soft computing [7][17][18][33][35][38] based Intrusion Detection Systems uses one or other form of data cleaning to provide quality data to detection engine and in turn results in improved detection rate. Often Intrusion Detection System plagued with huge amount of data to be processed. Processing this huge amount of data in real-time is another challenge faced by most Intrusion Detection Systems. Many researchers [21][31] have worked for increasing performance to give timely intrusion alerts. Reduction in input data rate will provide additional time to detection engine for thoroughly process data and give more detection accuracy with less false positive. Effective and more complex algorithms can be used for intrusion detection if sufficient time and resources available. To accomplish this, Intrusion Detection Systems use feature selection and extraction. Raw input data for intrusion detection consists of spurious and duplicate data. Not all the items available in raw input data are useful for intrusion detection. So we need to remove such spurious data which is not related to the purpose of intrusion detection. Another possible improvement used by modern Intrusion Detection Systems is removing duplicate content and generating summarized data which is directly useful for intrusion detection. Second functionality in our preprocessing module is feature selection and extraction. Removal of spurious and duplicate data will help in reducing false positive rate.

Another possible reason for false positive is lack of knowledge about network topology, live hosts and services running. In proposed model third functionality is configuration based processing. Configuration data about existing network, hosts, and services are stored in a file. Vulnerability assessment tools can also provide significant information. Specifically masquerader type of intrusion begins with IP scanning. Here ICMP ECHO request, ARP and other packets are used for live host detection. After identifying live host, masquerader needs to identify services running. Finally attacker uses exploitation scripts to exploit vulnerability. Configuration parameters help in differentiating normal and intrusion data by providing additional information. All Intrusion Detection Systems can easily identify certain data as intrusion while some other as attack. Some portion of overlapping behavior is the challenge for Intrusion Detection Systems. The data for which Intrusion Detection System is not sure results in false detection, either false negative or false positive. Such ambiguity can be reduced by collecting information from various sources. But this raises another challenge of dealing with multiple data formats. So in our proposed preprocessing module fourth functionality is to generate unified format from different data formats generated by different data sources. Using this unified log detection engine increase trust and can identify normal traffic. It will reduce chances of detecting as attack even though behavior is deviated from normal one. This again helps in reducing false positive rate in Intrusion Detection System.

4. RESULT

In our experiment raw network data is collected in the form of tcpdump. Vulnerability assessment is carried out with help of Linux BackTrack 5. Generating configuration information for preprocessing module is again manual process. Experimental result is highly dependent on this configuration data.

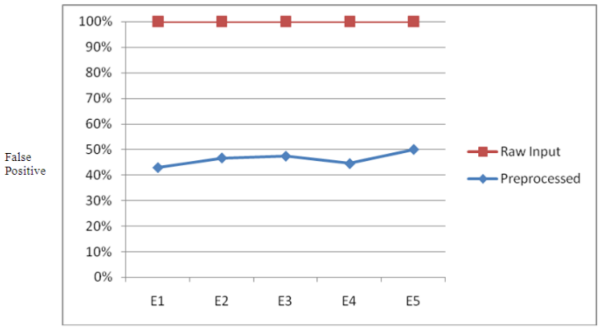

Figure 2: Preprocessed Vs Raw input data

Experimental results have clearly shown that even without fine tuning we can get reduction is false positive rate with the help of proposed preprocessing module. Base on the implementation of the preprocessing module and the target network environment results are changing but false positive reduction is observed.

5. CONCLUSION

In this paper, we have discussed preprocessing module as part of our proposed model for reducing false positive rate in Intrusion Detection System. We have proposed four major functionalities as part of preprocessing module. Removal of noise and incomplete data, feature selection and extraction, configuration parameter based processing, and unified log suggested for reduction of false positive rate. Our experimental results have shown that such preprocessing efforts can help Intrusion Detection System in reducing false positives. Efficient but more complex algorithms can be used for preprocessing in future. It will further help in reducing false positive but by consuming more resources.

6. REFERENCES

[1] A.A Abimbola, J.M Munoz and W.J Buchanan, Investigating False Positive Reduction in HTTP via Procedure Analysis, Proceedings of the International conference on Networking and Services, ISBN:0-7695-2622-5, Page 87, IEEE Computer Society Washington, DC, USA, July 2006

[2] Adelina Josephine D, Anushiadevi R, Lakshminarayanan T R, An Efficient Algorithm for Clustering Intrusion Alert, Journal of Theoretical and Applied Information Technology, Vol. 37 No.2, Pages 234-240, March 2012.

[3] Ala' Yaseen Ibrahim Shakhatreh, Kamalrulnizam Abu Bakar, A Review of Clustering Techniques Based on Machine learning Approach in Intrusion Detection Systems, IJCSI International Journal of Computer Science Issues, ISSN (Online): 1694-0814, Vol. 8, Issue 5, No 3, September 2011

[4] Amir Azimi Alasti Ahrabi, Ahmad Habibizad Navin, Hadi Bahrbegi, Mir Kamal Mirnia, Mehdi Bahrbegi, Elnaz Safarzadeh & Ali Ebrahimi, A New System for Clustering and Classification of Intrusion Detection System Alerts Using Self-Organizing Maps, International Journal of Computer Science and Security, (IJCSS), Volume (4): Issue (6)

[5] Aneetha.A.S., Revathi.S. and Bose.S, Dynamic Network Anomaly Intrusion Detection Using Modified SOM, Second International Conference on Computer Science, Engineering and Applications (CCSEA-2012), May 2012

[6] Ashley Thomas, RAPID: Reputation based approach for improving intrusion detection effectiveness, Sixth International Conference on Information Assurance and Security (IAS), Atlanta, GA, 2010

[7] B. Abdullah, I. Abd-alghafar, Gouda I. Salama and A. Abd-alhafez, Performance Evaluation of a Genetic Algorithm Based Approach to Network Intrusion Detection System, 13th International Conference on AEROSPACE SCIENCES & AVIATION TECHNOLOGY, ASAT- 13, May 26 – 28, 2009

[8] Bhawana Pillai, Uday Pratap Singh, NIDS for Unsupervised Authentication Records of KDD Dataset in MATLAB, (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Wireless & Mobile Networks, Page 57 – 61, ISSN 2156-5570 (Online), 2011

[9] Brian Eugene Lavender, Implementation of Genetic Algorithms into a Network Intrusion Detection System (netGA), and Integration into nProbe, M.S. Project, CALIFORNIA STATE UNIVERSITY, SACRAMENTO, 2010

[10] D.A. Karras, V. Zorkadis, Neural Network Techniques for Improved Intrusion Detection in Communication Systems, Proceedings of the 5th WSES International Conference on Circuits, Systems, Communications and Computers (CSCC 2001) ISBN: 960-8052-33-5, 2001

[11] Damiano Bolzoni, Sandro Etalle, APHRODITE: an Anomaly-based Architecture for False Positive Reduction, Cornell University Library, Subjects: Cryptography and Security (cs.CR), Report number: TR-CTIT-06-13, arXiv:cs/0604026v1 [cs.CR], 2006

[12] Deris tiawan, Abdul Hanan Abdullah, Mohd. Yazid dris, Characterizing Network Intrusion Prevention System, International Journal of Computer Applications (0975 – 8887), Volume 14– No.1, January 2011

[13] Dewan Md. Farid, Mohammad Zahidur Rahman, Chowdhury Mofizur Rahman, Adaptive Intrusion Detection based on Boosting and Naïve Bayesian Classifier, International Journal of Computer Applications (0975 – 8887), Volume 24– No.3, June 2011

[14] Dharmendra G. Bhatti, P. V. Virparia, Bankim Patel, Conceptual Framework for Soft Computing based Intrusion Detection to Reduce False Positive Rate, International Journal of Computer Applications (0975 – 8887), Volume 44– No13, April 2012

[15] Farhan Abdel-Fattah, Zulkhairi Md. Dahalin, Shaidah Jusoh, Distributed and Cooperative Hierarchical Intrusion Detection on MANETs, International Journal of Computer Applications (0975 – 8887) Volume 12– No.5, December 2010

[16] Fernando God´ınez and Dieter Hutter and Ra´ul Monroy, Service Discrimination and Audit File Reduction for Effective Intrusion Detection, Proceedings of the 5th international conference on Information Security Applications, Pages 99-113, Springer-Verlag Berlin, 2005

[17] G. Sunil Kumar, C.V.K Sirisha, Kanaka Durga.R, A.Devi, Robust Preprocessing and Random Forests Technique for Network Probe Anomaly Detection, International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, Volume-1, Issue-6, January 2012

[18] Gaurang Panchal, Parth Shah, Amit Ganatra , Y P Kosta, Unleashing Power of Artificial Intelligence for Network Intrusion Detection Problem, International Journal of Engineering Science and Technology, Vol. 2(10) , 5221-5230, 2010

[19] 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 2004 IEEE Workshop on Information Assurance and Security, United States Military Academy, West Point, NY, 10{11 June 2004

[20] Jing Xiao-Pei, Wang Hou-Xiang, A new Immunity Intrusion Detection Model Based on Genetic Algorithm and Vaccine Mechanism, I.J.Computer Network and Information Security, 2010, 2, 33-39

[21] Jintae Oh, Byoungkoo Kim, Seungyong Yoon, Jong-Soo Jang, Yong-Hee Jeon, and Jaecheol Ryou, A Novel Architecture and Mechanism for High-Performance Real-Time Intrusion Detection and Response System, International Journal of Computer Science and Network Security, VOL.8 No.3, March 2008

[22] Jiong Zhang and Mohammad Zulkernine, Anomaly based network intrusion detection with unsupervised outlier detection. In Symposium on network security and information assurance – proceedings of the IEEE international conference on communications (ICC), Istanbul, Turkey; June 2006

[23] Kai Hwang, Min Cai, Ying Chen, Min Qin, Hybrid Intrusion Detection with Weighted Signature Generation over Anomalous Internet Episodes, IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, VOL. 4, NO. 1, JANUARY-MARCH 2007

[24] KDDCUP 99 dataset, available at: http://kdd.ics.uci.edu/dataset/kddcup99/kddcup99.html

[25] M. Ali Aydin, A. Halim Zaim, K. Gökhan Ceylan, A hybrid intrusion detection system design for computer network security, Computers and Electrical Engineering, vol. 35, pp. 517-526, 2009.

[26] Mrutyunjaya Panda, Manas Ranjan Patra, Network Intrusion Detection Using Naive Bayes, IJCSNS International Journal of Computer Science and Network Security, Vol.7 No.12, December 2007

[27] Obbo Aggrey, An Intrusion Detection System For Academic Institutions, Master of Science Thesis, Makerere University, July 2007

[28] Priyanka Patil and Ujwala Patil, Preprocessing of web server log file for web mining, Proceedings of "National Conference on Emerging Trends in Computer Technology (NCETCT-2012), India, World Journal of Science and Technology, 2(3):14-18, ISSN: 2231 – 2587, 2012

[29] Rung-Ching Chen, Kai-Fan Cheng and Chia-Fen Hsieh, Using Rough Set and Support Vector Machine for Network Intrusion Detection, International Journal of

Network Security & Its Applications (IJNSA),Vol 1, No 1, April 2009

[30] Salem Benferhat, Karima Sedki, Karim Tabia, PREPROCESSING ROUGH NETWORK TRAFFIC FOR INTRUSION DETECTION PURPOSES, IADIS International Telecommunications, Networks and Systems 2007

[31] Sanjay Rawat, Arun K. Pujari, V. P. Gulati, On the Use of Singular Value Decomposition for a Fast Intrusion Detection System, Electronic Notes in Theoretical Computer Science 142 (2006) 215–228, Elsevier, 2006

[32] Shaimaa Ezzat Salama, Mohamed I. Marie, Laila M. El-Fangary & Yehia K. Helmy, Web Server Logs Preprocessing for Web Intrusion Detection, Computer and Information Science Vol. 4, No. 4; July 2011

[33] Shelly Xiaonan Wu, Wolfgang Banzhaf, The Use of Computational Intelligence in Intrusion Detection Systems: A Review, Technical Report #2008-05, Memorial University of Newfoundland, November 2008

[34] Sung-Bae Cho and Sang-Jun Han, Two Sophisticated Techniques to Improve HMM-Based Intrusion Detection Systems, Lecture Notes in Computer Science, Volume 2820/2003, SpringerLink, 2003

[35] Surat Srinoy, Werasak Kurutach, Witcha Chimphlee, and Siriporn Chimphlee, Network Anomaly Detection using Soft Computing, PROCEEDINGS OF WORLD ACADEMY OF SCIENCE, ENGINEERING AND TECHNOLOGY, VOLUME 9, ISSN 1307-6884, NOVEMBER 2005

[36] Tadeusz Pietraszek and Axel Tanner, Data Mining and Machine Learning - Towards Reducing False Positives in Intrusion Detection, Information Security Tech. Report (Elsevier Advanced Technology Publications Oxford, UK), Volume 10 Issue 3, Pages 169-183, January, 2005

[37] Witcha Chimphlee, Abdul Hanan Abdullah, Mohd Noor Md Sap, Siriporn Chimphlee, and Surat Srinoy, Unsupervised Clustering methods for Identifying Rare Events in Anomaly Detection, 6th Internation Enformatika Conference (IEC2005), October 26-28, Budapest, Hungary, 2005.

[38] Zheng Zhang, Jun Li, C.N. Manikopoulos, Jay Jorgenson, Jose Ucles, HIDE: a Hierarchical Network Intrusion Detection System Using Statistical Preprocessing and Neural Network Classification, Proceedings of the 2001 IEEE Workshop on Information Assurance and Security United States Military Academy, West Point, NY, 5-6 June, 2001

[39] Zorana Bankovic, Jose M. Moya, Alvaro Araujo, Slobodan Bojanic and Octavio Nieto-Taladriz, A Genetic Algorithm-based Solution for Intrusion Detection, Journal of Information Assurance and Security 4 (2009) 192-199, 2009