International Journal of Computer Networks and Communications Security

VOL. 4, NO. 12, DECEMBER 2016, 336–342
Available online at: www.ijcncs.org
E-ISSN 2308-9830 (Online) / ISSN 2410-0595 (Print)

Identifying the Most Frequently Attacked Ports Using Association Rule Mining

DOAA HASSAN

Computers and Systems Department, National Telecommunication Institute, Cairo, 11768, Egypt

doaa@nti.sci.eg

ABSTRACT

The security events presented in the dataset of network intrusion detection systems (NIDS) provide useful information about various network attacks that lunched against the network. For each security event, different information can be extracted including but not limited to the type of the event, the start and end time of the event, the source and destination IP addresses and ports. In this paper, we propose a novel framework for mining security events of DARPA-2009 network intrusion detection dataset. Our approach relies on finding a correlation between the security event of the dataset and the destination port that is exploited by an attacker in order to hack the network according to what security event reports. Association rule mining technique has been used in this paper to discover such correlation, since it is widely used to find strong correlations between features of massive datasets in terms of generated rules. Thus, the proposed framework aims to discover the most frequently destination ports that often exploited by an attacker to lunch a network attack according to the generated rules. This can save the time required to manually categorize destination ports in view of the security events reported in the dataset. Moreover, it can be very useful in creating various security policies for blocking the ports that can be used as back-doors by the attacker that let him/her illegally accesses the network. Various sets of rules have been generated using association rules based on apriori algorithm for the experimental analysis of the proposed approach. The performance of this algorithm is compared based on various interestingness measures.

Keywords: *Network Intrusion Detection, Association Rule Mining, Apriori Algorithm, Port Attack Detection.*

1 INTRODUCTION

Recently, the great expansion of networks and their users has led to the creation of new security threats against networks. Such threats can be caused by many hacking tools and intrusive methods scenarios. As a result, network intrusion detection systems (NIDS) were introduced to detect the abnormal activities within a network, expressed in a form of a list of security alerts [18]. NIDS is deployed as an effective hardware [19] or software [20] that can be installed at the edge of large enterprise networks in order to identify the pattern of intrusions caused by various types of attacks lunched by an attacker against the network.

For NIDS evaluation, there has been various datasets that have been widely used in the literature such as KDD cup99 [22], NSLKDD [23], DARPA-2009 datasets [9, 10]. These datasets have a large collection of normal and anomaly network traces with different type of features that address different parameters for each network trace.

Data mining has been extensively known as a common research technique that utilizes NIDS datasets for extracting implicit signs of intrusions from those datasets and enables transforming those singes to users [21]. In general, using data mining aims to achieve two goals, namely, description and prediction [11]. Description aims to find some human meaningful patterns that describe the data such as grouping or clustering techniques. Prediction uses some features of the dataset to predict unknown values of other features of interest. Classification is an example of prediction.

The proposed framework in this paper does not predict new threats. Instead, it aims to detect the most frequently dangerous ports through which various attack can be launched. Therefore, the proposed framework represents the relationship between the attacked/dangerous ports and the type of attack reported in the security events report of

NIDS dataset. Such relationship is identified using association rule mining; a common data mining technique for finding the strong correlations between features of massive datasets in terms of generated rules. The extraction of such relationship can save the time required to manually categorize destination ports for determining the dangerous ones in view of the security events of NIDS dataset. Moreover, it can be very useful in creating various security policies for blocking the ports that can be used as back-doors by the attacker that let him/her illegally accesses the network.

The experimental analysis of the proposed approach takes several steps. First the security events ground truth data of DARPA-2009 dataset [10] has been used for the analysis. The choice of DARPA-2009 dataset is due to that it has more comprehensive reflection for the recent network threat environment than KDD-cup 99 and NSL-KDD datasets that have been used extensively addressed in NIDS literature. Moreover, DARPA-2009 has a specific attribute for destination ports, which can help in the analysis required for port scan attack detection [8]. Second, various sets of rules have been generated using association rules based on apriori algorithm [24] to address the relation between attack types and destination ports. Finally, the performance of apriori algorithm for rule generation is compared based on various interestingness measures.

The rest of this paper is organized as follows: In Section 2, we provide a background of association rule mining and an overview of the security events of DARPA-2009 dataset. In Section 3, we introduce our proposed approach as well as the experimental evaluation. In Section 4, we discuss the related work. Finally we conclude the paper in Section 5 with a direction for future work.

2 BACKGROUND

2.1 Overview of Association Rules Mining

Association rule mining is an important technique of data mining for discovering hidden interesting relationships between features in massive datasets [11]. Such relationship can be represented by a set of strong rules generated from the dataset based on some measures of interestingness. Each rule found in the dataset has following form:

$$A \Rightarrow B \qquad (1)$$

where A is called antecedent and B is called consequent. This rule suggests an interesting relationship between A and B, where A and B are two sets of items called itemsets that are subsets of a set of n binary attributes called items I= {i_1; i_2; i_3; i_n} and A∩ B = ϕ; (i.e., A and B do not have any items in common). The dataset from which the rules are generated has a set of transactions. Each transaction has an ID and a set of items that is a subset of I.

Two widely used measures of interestingness, namely a minimum threshold on the support and confidence are used for selecting interesting rules from the set of all possible generated rules. Support is an indication of how often item-set appears in the dataset. For example the support for item-set A in (1) is given by:

$$support(A) = \frac{support - count(A)}{N} \qquad (2)$$

where support-count(A) is the number of times that A appears in the dataset and N is the number of transactions in the dataset. Confidence is an indication of how often the rule has been found to be true. For example, if we consider the rule in (1), then the confidence of the rule refers to how many times the items of B appear in a transaction that contains A. Formally this can be expressed as follows:

$$confidence(A \Rightarrow B) = \frac{support(A \cup B)}{support(A)} \qquad (3)$$

There other three common measures that are also used to evaluate the interestingness of the rule, namely lift, conviction and leverage. Lift is the ratio between the rule confidence and the support of the rule consequent. Formally this can be expressed as follows:

$$lift(A \Rightarrow B) = \frac{confidence(A \Rightarrow B)}{support(B)} \qquad (4)$$

Conviction of a rule is defined formally as:

$$conv(A \Rightarrow B) = \frac{1 - support(B)}{1 - confidence(A \Rightarrow B)} \qquad (5)$$

Finally, leverage of the rule is defined as:

$$lev(A \Rightarrow B) = support(A \cup B) - support(A) * support(B) \qquad (6)$$

There are common three existing algorithms used for association rule mining including apriori, predictive apriori, and Tertius algorithms [25]. In this paper, we use apriori algorithm [24] for generating attack type-port association rules that identify the hidden relationship between attack-types and destination ports. The choice of apriori algorithm was due that it is the best known algorithm for mining association rule as it has an innovative way for generating association rules on large scale datasets. The following are the two basic principles of the algorithm:

- All subsets derived from a frequent itemset are also frequent.
- All subsets derived from infrequent itemset are also infrequent.

where the infrequent itemset is distinguished from the frequent one according to its support value which is less than the support value of the frequent itemset. Apriori uses pruning based on the support value to stop the exponential growth of the generated candidate itemset, where every item is initially considered as a candidate-1 itemset and the candidate itemsets that have a support value that is less than the minimum support value are discarded.

2.2 Security Events of DARPA 2009 Dataset

The DARPA-2009 dataset [9] is a rich dataset that contains recent attack vectors. The DARPA 2009 dataset is created with synthesized traffic to emulate traffic between a /16 subnet (172.28.0.0/16) and the Internet. The traces in the dataset have been captured in 10 days between the 3rd and the 12th of November of the year 2009. They represent various types of data traffic such as DNS, synthetic HTTP, and SMTP traffic. The dataset has a list of 46 security events that are documented through the 10 days in the security event report of DARPA-2009 dataset [10]. Those events include modern styles of attack types and worms such as different distributed denial of service (DDoS) attacks, Spambots, and Phishing Emails. In this paper, the ground-truth data for the security events [13] is used as the experimental dataset. The basic information about the events is reported in this ground-truth dataset. This includes the event type, source and destination IPs and ports, the start and end time of the event. We refer to [10, 6] for more information about DARPA-2009 dataset and its security events.

3 PROPOSED APPROACH

The proposed approach uses the DARPA-2009 security events ground-truth dataset [13] that consists of a collection of 8223 security events that represent different attack types. Such collection is resulting from capturing and analyzing network traces between an internal network and the internet in 10 days. All security events in the dataset are summarized in [14]. The proposed approach has two main phases, namely the data-preprocessing phase and generation of attack-type port rules phase. Figure 1 illustrates our proposed approach.

3.1 Data-Preprocessing

The first phase in our approach as shown in figure 1 is to preprocess this dataset and perform data cleaning in order to extract fields with only information about attack types and destination ports for constructing apriori training dataset. This dataset is used to create a model that computes the probability of either using a specific port for initializing one or more type of network attacks, or lunching a certain network attack through various destination ports. Thus, the apriori algorithm creates a model to detect attack types that have been lunched from a specific port or detecting destination ports that have been used by an attacker for lunching a particular network attack. This will lead to extracting the most frequently attacked ports.

3.2 Attack Types-Port Rules Generation

The second phase in the proposed approach is to use the association rule mining based on apriori algorithm for generating attack types-port rules. Those rules are generated by finding the interestingness relationship between attack types and destination ports; hence we can identify the most frequent ports that are exploited by an attacker to lunch a network attack. In summary, the attack types- port rules are generated by the following two steps:

- Mining the security events according to destination ports or vice versa. From security events, one can find sets of security events (i.e., network attacks) that are commonly lunched from a specific destination port, while from destination ports; one can find the sets of destination ports that are commonly used to lunch a specific network attack

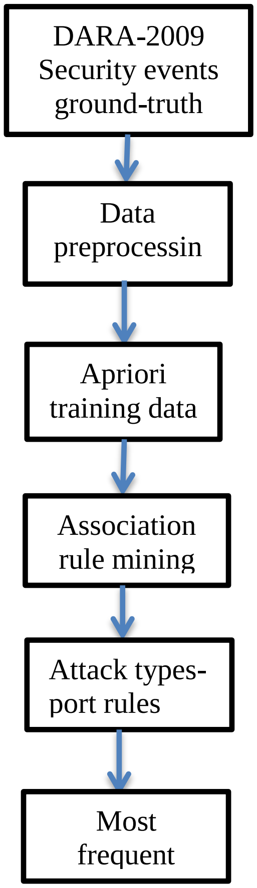

Fig. 1. Proposed approach.

- Extracting interesting rules in a form of implications between set of security events and a destination port or vice versa, with a confidence and support that exceed the minimum confidence and support specified by user. An example of such rule is as follows:

Event Type=failed attack exploit/iis-asp-overflow 242==>Destination Port(s)=80.0 242 *conf:(1)*

The rule shows an interestingness relationship between a set of two security events, namely the failed attack exploit and IIs-asp-overflow and port 80 (i.e., the HTTP port). This indicates that all failed attack exploit, reported in the security events ground-truth data is lunched from port 80. Moreover, the attacker usually lunches IIS-asp overflow attack through the same port. The rule support is 242 and its confidence is 1, which means the support account for antecedent of the rule is equal to the support count for the consequent of the rule.

3.3 Experimental Evaluation

We have conducted three experiments on the preprocessed dataset. In each experiment, we have selected different value for the minimum support. Various interestingness measures are used to evaluate the generated rules including: the confidence, lift, conviction and leverage. Thus in the end we have carried out 12 experiment, where one of the interestingness measures has been used each time with three different values for the minimum support. The experiments have been conducted on a windows 7 laptop machine with 2.6 GHZ processor Intel core (TM) i5 and 4 G Memory Rams. Weka [12], a free data mining software tool has been used for generating attack types-port rules using association rule mining based on apriori algorithm [24]. Table 1 shows the mining results corresponding to each value of minimum support, when considering confidence as the interestingness measures.

As concluded from the results shown in the table, the most frequently attacked ports or those that are used to lunch a specific network attack are {25, 80, 499, 1257, 3128,10000}.

4 RELATED WORK

Association rule mining for network intrusion detection: There has been some research work that used association rule mining for finding correlation between features of intrusion detection datasets such as those presented in [1, 4, 7].

For example, the work presented in [1] proposed a model that uses association rule based on Apriori algorithm for generating real-time rules for firewall in order to predict anomaly attacks. The model used Snort; a network-based intrusion detection software to record logs of user activities, and then apriori algorithm was used to generate online rules for firewall based on recorded user activities.

In [4], association rule mining was used to determine the correlation between various attributes of the KDD-cup99 network intrusion detection dataset and attack type attribute. Such a correlation was expressed by generating

Table 1: The attack type-port rules generated for three different values of minimum support with confidence interestingness measure

Minsup value	Generated rules
0.1	Event Type=ddos 2983 ==> Destination Port(s)=80.0 2982 conf:(1)
0.01	.1Event Type=failed attack exploit/iis-asp-overflow 245 ==> Destination Port(s)=80.0 245 conf:(1) .2 Destination Port(s)=1257 3128 234 ==> Event Type=scan /usr/bin/nmap 234 conf:(1) .3 Event Type=phishing email exploit/malware/trawler 231 ==> Destination Port(s)=25.0 231 conf:(1) .4 Destination Port(s)=499.0 161 ==> Event Type=malware ddos 161 conf:(1) .5 Event Type=malware ddos 161 ==> Destination Port(s)=499.0 161 conf:(1) .6 Event Type=noisy c2+ tcp control channel exfil nc 150 ==> Destination Port(s)=10000.0 150 conf:(1) .7 Event Type=post-phishing client compromise + malicious download 126 ==> Destination Port(s)=80.0 126 conf:(1) .8 Event Type=ddos 2983 ==> Destination Port(s)=80.0 2982 conf:(1)
0.001	1. Event Type=failed attack exploit/iis-asp-overflow 245 ==> Destination Port(s)=80.0 245 conf:(1) 2. Destination Port(s)=1257 3128 234 ==> Event Type=scan /usr/bin/nmap 234 conf:(1) 3. Event Type=phishing email exploit/malware/trawler 231 ==> Destination Port(s)=25.0 231 conf:(1) 4. Destination Port(s)=499.0 161 ==> Event Type=malware ddos 161 conf:(1) 5. Event Type=malware ddos 161 ==> Destination Port(s)=499.0 161 conf:(1) 6. Event Type=noisy c2+ tcp control channel exfil nc 150 ==> Destination Port(s)=10000.0 150 conf:(1) 7. Event Type=post-phishing client compromise + malicious download 126 ==> Destination Port(s)=80.0 126 conf:(1) 8. Event Type=no precursor client compromise exfil/sams_launch_v 58 ==> Destination Port(s)=80.0 58 conf:(1) 9. Event Type=noisy c2+ tcp control channel exfil- fork 50 ==> Destination Port(s)=10000.0 50 conf:(1) 10. Event Type=c2 remote command execution nc 48 ==> Destination Port(s)=10000.0 48 conf:(1)

rules using apriori algorithm, where each rule takes the form of if-then-else structure such as if a specific attribute takes a certain value, then a certain type of attack occurs. Those rules were used as a basis to test if the generated rule can detect the intrusion on the test set.

The work presented in [7] compared the classification association rules (CARs) that were generated using apriori algorithm and predictive apriori algorithm. Those rules were generated by finding the correlation between attributes of either Snort logs or KDD-cup99 dataset in order to distinguish between the relevant and irrelevant alerts, and hence discovering the most critical threads.

Port Mining and discovering relation with attack types: The relation between the information available about destination ports and the type of attack has been studied in some research work such as what was presented in [15-17, 2].

Ban et el. [2] presented a model that uses the linkage algorithm for grouping/clustering hosts that attack an almost identical list of destination ports. The dissimilarity of two hosts is measured using Jaccard distance, where two hosts are clustered together if their Jaccard distance is less than a cutoff parameter c.

Cheng and Y. Tang [3] presented PortView, a model for identifying port roles based on port fuzzy macroscopic behavior which can detect new behavior of network attack. The port roles are determined by identifying the association between port by its number and application types. Their model uses an EM fuzzy clustering algorithm in order to classify port roles into six role categories including the scanned ports, failed service ports, client ports, server ports, P2P ports, and overlay ports. They used real network traffic captured from a link in the CERNET network in order to validate their model.

The closest work to ours is the one proposed by Al-Saedi et el. [5]. They presented an algorithm for mining network-worm affinities, where affinities are mined in two steps: mining port group that were attacked by common group of worms. Second, using the threat degree as a factor for extracting the association of threatening affinities in a form of implications that are grouped according to a user-defined threshold. Our work is similar to their work in that it also mines port groups that were attacked by common attack. However, our work also considers the inverse case, by mining the attack (i.e., security events) groups that were lunched from a common port. Also we are not concerned with finding implications/associations between network-worm affinities.

5 CONCLUSIONS AND FUTURE WORK

This paper has proposed a model that identifies the most frequently ports that are exploited by an attacker to launch a network attack. The model uses association rule mining for generating rules that address this issue. The rules have been generated by selecting two relevant features from a security report of DARPA-2009 dataset, namely the event type and the destination port. DARPA-2009 ground-truth data has been used as it records a large collection of security events that consists of enormous attack types. Apriori algorithm has been used to create a model that generates attack types-port rules by identifying interestingness relationship between attack types and destination ports. Various interesting measures have been used to evaluate the performance of this algorithm for generating strong and interesting rules. The proposed model can be very useful in automatically categorizing destination ports in view of the security events, and hence blocking the dangerous ports that could be used as a backdoor by the attacker that enable him/her to hack the network.

As a future work, we are looking forward to testing our proposed approach for generating attack type-port association rules using other algorithms and then comparing them with those generated using apriori algorithms. Also, we are looking forward to applying our proposed approach to other real network intrusion detection datasets.

6 REFERENCES

[1] E. Saboori, S. Parsazad and Y. Sanatkhani. Automatic Firewall rules generator for Anomaly Detection Systems with Apriori algorithm. In Proceedings of the 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), Volume 6, pp.57-60, 2010.

[2] T. Ban, L. Zhu, J. Shimamura, S. Pang, D. Inoue, and K. Nakao. Behavior Analysis of Long-term Cyber Attacks in the Darknet. In Proceedings of the 19th International Conference on Neural Information Processing (ICONIP 2012), Doha, Qatar, November 12-15, 2012.

[3] G. Cheng and Y. Tang PortView: identifying port roles based on port fuzzy macroscopic behavior. Journal of Internet Services and Applications, Volume 4, No. 1, pp9:1-9:12, 2013.

[4] F. S. Tsai Network Intrusion Detection Using Association Rules. International Journal of Recent Trends in Engineering, Vol 2, No. 2, November 2009.

[5] K. H. Al-Saedi, H. Al-Khafaji and A. ALmomani, S. Manickam, and Sureswaran Ramadass. An Approach to Assessment of Network Worm Detection Using Threatening-Database Mining. Australian Journal of Basic and Applied Sciences, Volume 5, No. (12), pp-2676-2683, 2011.

[6] N. Moustafa and J. Slay Creating Novel Features to Anomaly Network Detection Using DARPA-2009 Data set In of Proceedings of the 14th European Conference on Cyber Warfare and Security, 2015.

[7] B. T. Saputra and F. Yang. Improving IDS Alerts Using Predictive Apriori Algorithm. In Proceeding of International Multi-Conference on Engineering and Tech-Innovation 2015 (IMETI2015), October 30-November 03, 2015, Kaohsiung, Taiwan.

[8] C. Bailey, L. Chris Roedel and E. Silenok. Detection and Characterization of Port Scan Attacks. Available at: cseweb.ucsd.edu/~clbailey/PortScans.pdf.

[9] DARPA 2009 Intrusion Detection Dataset. Available at: http://www.darpa2009.netsec.colostate.edu/.

[10] M. Gharaibeh and C. Papadopoulos. DARPA-2009 Intrusion Detection Dataset Report. Available at: http://www.darpa2009.netsec.colostate.edu/DARPA_Set_ Report.pdf, 2014.

[11] P-N. Tan, M. Steinbach, and V. Kumar. Introduction to Data Mining, 1st Edition, Addison-Wesley, 2005.

[12] E. Frank, Mark A. Hall, and Ian H. Witten. The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

[13] Security events ground-truth data. Available at: http://www.darpa2009.netsec.colostate.edu/DARPA_Groundtruth_10_day_test.xlsx.

[14] List of dataset security events. Available at: http://www.darpa2009.netsec..edu/List_of_security_events.pdf

[15] J. Vinu and T. Theepak. Realization of comprehensive botnet inquisitive actions. International Conference on Computing, Electronics and Electrical Technologies (ICCEET), pp. 915921, March 2012.

[16] K. Limthong, F. Kensuke, P. Watanapongse. Wavelet-based unwanted traffic time series analysis. In proceedings of International Conference on Computer and Electrical

Engineering, (ICCEE 2008), pp. 445449, December, 2008.

[17] C. McManamon, and F. Mtenzi. Defending privacy: The development and deployment of a darknet. In: 2010 International Conference for Internet Technology and Secured Transactions (ICITST), pp. 16, Novemeber, 2010.

[18] Lata and I. Kashyap Study and Analysis of Network based Intrusion Detection System. International Journal of Advanced Research in Computer and Communication Engineering,Vol.2, Issue5, May, 2013.

[19] C. Clark, W. Lee, D. Schimmel, D. Contis, M. Kon, and A. Thomas. A Hardware Platform for Network Intrusion Detection and Prevention. In Proceedings of Workshop on Network Processors and Applications (NP3), pp. 136-145, 2004.

[20] R. H. Gong, M. Zulkernine, P. Abolmaesumi. A Software Implementation of a Genetic Algorithm Based Approach to Network Intrusion Detection. Proceedings of the Sixth International Conference on Software Engineering, Arti_cial Intelligence, Networking and Parallel/Dis tributed Computing and First ACIS International Workshop on Self-Assembling Wireless Networks (SNPD/SAWN05), 2005.

[21] E. Bloedorn, A. D. Christiansen, W. Hill, C. Skorupka, L. M. Talbot, J. Tivel. Data Mining for Network Intrusion Detection: How to Get Started MITRE, Technical Report, August, 2001.

[22] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani. A Detailed Analysis of the KDD CUP 99 Data Set. In Proceedings of the IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA 2009), 2009.

[23] S. Revath and Dr. A. Malathi. A Detailed Analysis on NSL-KDD Dataset. Using Various Machine Learning Techniques for Intrusion Detection. International Journal of Engineering Research & Technology (IJERT), Vol. 2, Issue 12, December 2013.

[24] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association. In Proceedings of the 20th VLDB Conference, San tiago, Chile, 1994.

[25] Mukesh Sharma, Jyoti Choudhary, and Gunjan Sharma Evaluating the performance of apriori and predictive apriori algorithm to _nd new association rules based on the statistical measures of datasets. International Journal of Engineering Re- search & Technology (IJERT)),Vol. 1, Issue 6, August, 2012.

AUTHOR PROFILES:

Dr. Doaa Hassan earned her Ph.D. in January, 2012 from Computer and Systems Engineering Department at Faculty of Engineering at Zagazig University - Egypt. As a part of her PhD, she also spent one year and half as Ph.D. candidate at Computer Science Department at Eindhoven University of Technology in Netherlands. Currently, she is affiliated as an Assistant Professor at Computers and Systems Department at National Telecommunication Institute in Cairo- Egypt. She is also a visiting research associate at School of Informatics and Computing at Indiana University - Bloomington. Her research interest focuses on using machine learning and data mining techniques for automatic detection of network intrusions and applications malware and enforcement of information flow policies.