

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900

Open access books available

185,000

International authors and editors

200M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



Introductory Chapter: Machine Learning in Misuse and Anomaly Detection

Jaydip Sen and Sidra Mehtab

1. Introduction

Over the last 30 years, ubiquitous and networked computing has increasingly gained importance in our life. With the increase in complexity of computer networks, cybersecurity threats have also manifested in a variety of which was unimaginable even a decade back. While the rule-based intrusion detection systems (IDSs) can accurately detect already known attacks on a cyberinfrastructure, these systems are not capable to detect novel, unknown, and polymorphic cyber threats. Moreover, the computational overheads including CPU cycles and memory overheads are unacceptably high for most of the detection systems. Hence, it has been a constant challenge for security researchers to design automated, fast, and yet accurate IDSs for deployment in real-world cyberinfrastructures. From expert-crafted rules to sophisticated machine learning and deep learning algorithms, researchers have explored and attempted to push the boundary of the detection accuracy while minimizing the false alarm rates.

Applications of machine learning and data mining algorithms in both signature and anomaly detection systems have been widely proposed in the literature. In misuse detection systems, following approaches of machine learning are quite popular: (1) classification using association rules [1–3], (2) artificial neural networks [4], (3) support vector machines [5], (4) classification and regression trees [6, 7], (5) Bayesian network classifier [8–10], and (6) naïve Bayes method [11]. While the signature detection systems require labeled training data in order to learn the features of the attack and the normal traffic, anomaly detection systems are based on identifying any significant changes in the system from its normal state. Various approaches to machine learning in anomaly detection have been proposed in the literature. Some of these approaches are as follows: (1) association rule mining [12–14], (2) fuzzy association rule mining [15], (3) artificial neural network [16–18], (4) support vector machines [19, 20], (5) nearest neighbor [21], (6) hidden Markov model [22–24], (7) Kalman filter [25], (8) clustering [26], and (9) random forest [27, 28]. Other machine learning methods have been proposed for learning the probability distribution of data and in applying statistical tests to detect outliers [29–35].

The hybrid detection approach combines the adaptability and the powerful detection ability of an anomaly detection system with the higher accuracy and reliability of the misuse detection approach [28, 36–43]. The selection of misuse and anomaly detection systems for designing a hybrid detection system is dependent on the application in which the detection system is to be deployed. Following a combinational approach, the integration of an anomaly detection system with a misuse detection counterpart has been classified into four categories [28, 36]. These types

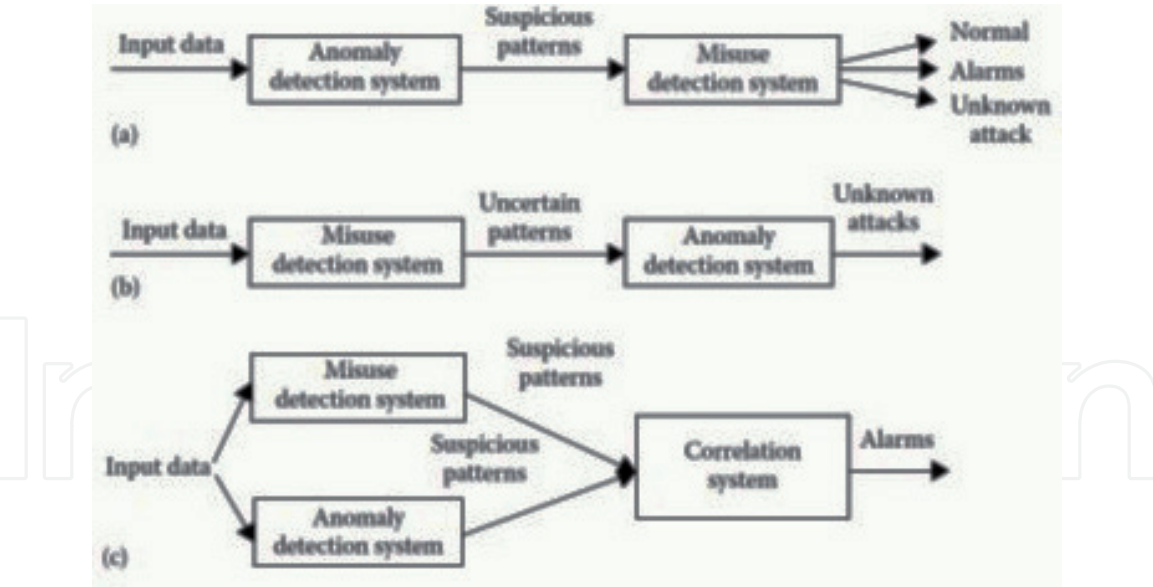


Figure 1. Three categories of hybrid detection systems. (a) Anomaly–misuse sequence, (b) misuse–anomaly sequence, and (c) parallel detection system.

are: (1) anomaly–misuse sequence detection, (2) misuse–anomaly sequence detection, (3) parallel detection, and (4) complex mixture detection. The complex mixture model is highly application-specific. **Figure 1** depicts the first three categories of hybrid detection systems. Complex detection systems are application-specific, and these systems cannot be represented by any generic architecture.

2. Conclusion

A fundamental challenge in designing an intrusion detection system is the limited availability of appropriate data for model building and testing. Generating data for intrusion detection is an extremely painstaking and complex task that mandates the generation of normal system data as well as anomalous and attack data. If a real-world network environment, generating normal traffic data is not a problem. However, the data may too privacy-sensitive to be made available for public research.

Classification-based methods require training data to be well balanced with normal traffic data and attack traffic data. Although it is desirable to have a good mix of a large variety of attack traffic data (including some novel attacks), it may not be feasible in practice. Moreover, the labeling of data is mandatory with attack and normal traffic data clearly distinguished by their respective labels.

Unlike classification-based approaches that are mostly used in misuse detection, unsupervised anomaly detection-based approaches do not require any prior labeling of the training data. In most of the cases, the attack traffic constitutes the sparse class, and hence, the smaller clusters are most likely to correspond to the attack traffic data. Although unsupervised anomaly detection is a very interesting approach, the results produced by this method are unacceptably low in terms of their detection accuracies.

In a pure anomaly detection approach, the training data are assumed to be consisting of only normal traffic. By training the detection model only on the normal traffic data, the detection accuracy of the system can be significantly improved. Anomalous states are indicated by only a significant state change from the normal state of the system.

In a real-world network that is connected to the Internet, an assumption of attack free traffic is utopian. A pure anomaly detection system can still be trained on a training data that include attack traffic. In that case, those attack traffic data will be considered as normal traffic, and the detection system will not raise an alert when such traffic is encountered in real-world operations. Hence, in order to increase the detection accuracy, attack traffic should be removed from the training data as much as possible. The removal of attack traffic from the training data can be done using updated misuse detection systems or by deploying multiple anomaly detection systems and combining their results by a voting mechanism.

For an intrusion detection system that is deployed in a real-world network, it is mandatory to have a real-time detection capability under a high-speed, high-volume data environment. However, most of the cluster techniques used in unsupervised detection require quadratic time. This renders their deployment infeasible in practical applications. Moreover, the cluster algorithms are not scalable, and they need the entire training data to reside in the memory during the training process. This requirement puts a restriction on the model size. The future direction of research may include studies on scalability and performance of anomaly detection algorithms in conjunction with the detection rate and false positive rate. Most of the currently existing propositions on intrusion detection have not paid adequate attention to these critical issues.


In this book, the following chapters deal with various aspects of network security and cryptography. While the chapters belonging to the network security section broadly discuss different aspects of applications and deployment of security protocols and secure system architecture, the cryptography section discusses various theoretical algorithms and their complexities.

Author details

Jaydip Sen* and Sidra Mehtab
School of Computing and Analytics, NSHM Knowledge Campus, Kolkata, India

*Address all correspondence to: jaydip.sen@acm.org

IntechOpen

© 2020 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. In: Proceedings of the ACM SIGMOD International Conference on Management of Data. Washington, DC: ACM; 1993. pp. 207-216
- [2] Lee WK, Stolfo SJ, Mok KW. A data mining framework for building intrusion detection models. In: Proceedings of the IEEE Symposium on Security and Privacy. Oakland, CA: IEEE; 1999. pp. 120-132. DOI: 10.1109/SECPRI.1999.766909
- [3] Abraham A, Grosan C, Martin-Vide C. Evolutionary design of intrusion detection programs. International Journal of Network Security. 2007;4(3):328-339. DOI: 10.6633/IJNS.200705.4(3).12
- [4] Cannady J. Artificial neural networks for misuse detection. In: Proceedings of the National Information Systems Security Conference (NISSC'98); Washington, DC; 6-9 October 1998. pp. 441-454
- [5] Mukkamala S, Janoski G, Sung AH. Intrusion detection using neural networks and support vector machines. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN'02); Honolulu, HI; 12-17 May 2002. pp. 1702-1707. DOI: 10.1109/IJCNN.2002.1007774
- [6] Kruegel C, Toth T. Using detection trees to improve signature-based intrusion detection. In: Proceedings of the 6th International Workshop on Recent Advances in Intrusion Detection; Pittsburgh, PA; 8-10 September 2003. pp. 173-191. DOI: 10.1007/978-3-540-45248-5_10
- [7] Chebrolu S, Abraham A, Thomas JP. Feature deduction of intrusion detection systems. Computers & Security. 2005;24:295-307. DOI: 10.1016/j.cose.2004.09.008
- [8] Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Machine Learning. 1992;9:309-347. DOI: 10.1007/BF00994110
- [9] Verma T, Pearl J. An algorithm for deciding if a set of observed independencies has a causal explanation. In: Proceedings of the 8th International Conference on Uncertainty in Artificial Intelligence; Stanford, CA; 1992. pp. 323-330. DOI: 10.1016/B978-1-4832-8287-9.50049-9
- [10] Pearl J, Wermuth N. When can association graphs admit a causal interpretation? In: Proceedings of the 4th International Workshop on Artificial Intelligence and Statistics; Fort Lauderdale, FL; 1993. pp. 141-150. DOI: 10.1007/978-1-4612-2660-4_21
- [11] Schultz MG, Eskin E, Zadok E, Stolfo SJ. Data mining methods for detection of new malicious executables. In: Proceedings of IEEE Symposium on Security and Privacy (S&P'01); Oakland, CA/Anaheim, CA; 14-16 May 2000. DOI: 10.1109/SECPRI.2001.924286
- [12] Lee W, Stolfo SJ. Data mining approaches for intrusion detection. In: Proceedings of the 7th USENIX Security Symposium; San Antonio, TX; 26-29 January 1998. DOI:10.7916/D86D60P8
- [13] Apiletti D, Baralis E, Cerquitelli T, D'Elia V. Characterizing network traffic by means of the NetMine framework. Computer Networks. 2009;53(6):774-789. DOI: 10.1016/j.comnet.2008.12.011
- [14] Mannila H, Toivonen H. Discovering generalized episodes using minimal occurrences. In: Proceedings of the 2nd International Conference on

Knowledge Discovery in Databases and Data Mining. Portland, OR: ACM; 1996. pp. 146-151

[15] Luo J, Bridges SM. Mining fuzzy association rules and fuzzy frequency episodes for intrusion detection. *International Journal of Intelligent Systems*. 2000;15(8):687-703

[16] Ghosh AK, Wanken J, Charron F. Detecting anomalous and unknown intrusions against programs. In: *Proceedings of the 14th Annual Computer Security Applications Conference (ACSAC'98)*; Phoenix, AZ; 7-10 December 1998. DOI: 10.1109/CSAC.1998.738646

[17] Ghosh AK, Schwartzbard A, Schatz M. Learning program behavior profiles for intrusion detection. In: *Proceedings of the 1st USENIX Workshop on Intrusion Detection and Network Monitoring*; Santa Clara, CL; 9-12 April 1999. pp. 51-62

[18] Liu Z, Florez G, Bridges SM. A comparison of input representations in neural networks: A case study in intrusion detection. In: *Proceedings of the International Joint Conference on Neural Networks (IJCNN'02)*; Honolulu, HI; 12-17 May 2002. DOI: 10.1109/IJCNN.2002.1007775

[19] Chen WH, Hsu SH, Shen HP. Application of SVM and ANN for intrusion detection. *Computers and Operations Research*. 2005;32(10):2617-2634. DOI: 10.1016/j.cor.2004.03.019

[20] Hu WJ, Liao YH, Vemuri VR. Robust support vector machines for anomaly detection in computer security. In: *Proceedings of the International Conference on Machine Learning (ICMLA'03)*; Los Angeles, CL: CSREA; 23-24 June 2003. pp. 161-167

[21] Liao YH, Vemuri VR. Use of k -nearest neighbor classifier for intrusion detection. *Computers &*

Security. 2002;21(5):439-448. DOI: 10.1016/S0167-4048(02)00514-X

[22] Warrender C, Forrest S, Pearlmuter B. Detecting intrusions using system calls: Alternative data models. In: *Proceedings of IEEE Symposium on Security and Privacy*; Oakland, CA: IEEE; 10-14 May 1999. pp. 133-145. DOI: 10.1109/SECPRI.1999.766910

[23] Qiao Y, Xin XW, Bin Y, Ge S. Anomaly intrusion detection method based on HMM. *Electronics Letters*. 2002;38(13):663-664. DOI: 10.1049/el:20020467

[24] Wang W, Guan X, Zhang X, Yang L. Profiling program behavior for anomaly intrusion detection based on the transition and frequency property of computer audit data. *Computers & Security*. 2006;25(7):539-550. DOI: 10.1016/j.cose.2006.05.005

[25] Soule K, Salamatian K, Taft N. Combining filtering and statistical methods for anomaly detection. In: *Proceedings of the 5th ACM SIGCOMM Conference on Internet Measurement*. Berkeley, CA: ACM; 19-21 October 2005. pp. 331-344. DOI: 10.1145/1330107.1330147

[26] Portnoy L, Eskin E, Stolfo S. Intrusion detection with unlabeled data using clustering. In: *Proceedings of ACM CSS Workshop on Data Mining Applied to Security (DMSA)*. Philadelphia, PA: ACM; 2001. pp. 5-8. DOI: 10.7916/D8MP5904

[27] Zhang J, Zulkernine M. Anomaly based network intrusion detection with unsupervised outlier detection. In: *IEEE International Conference on Communications*. Istanbul, Turkey: IEEE; 11-15 June 2006. pp. 2388-2393. DOI: 10.1109/ICC.2006.255127

[28] Zhang J, Zulkernine M, Haque A. Random forest-based network intrusion detection systems. *IEEE Transactions*

on Systems, Man, and Cybernetics, Part C: Applications and Reviews. 2008;**38**(5):649-659. DOI: 10.1109/TSMCC.2008.923876

[29] Eskin E. Anomaly detection over noisy data using learned probability distribution. In: Proceedings of the 17th International Conference on Machine Learning (ICML'00). Stanford, CA: ACM; 29 June–2 July 2000. pp. 255-262. DOI: 10.7916/D8C53SKF

[30] Ye N, Li X, Chen Q, Emran SM, Xu M. Probabilistic techniques for intrusion detection based on computer audit data. IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans. 2001;**31**(4):266-274. DOI: 10.1109/3468.935043

[31] Feinstein L, Schnackenberg D, Balupari R, Kindred D. Statistical approaches to DDoS attack detection and response. In: Proceedings of DARPA Information Survivability Conference and Exposition. Washington, DC: IEEE; 2003. pp. 303-314. DOI: 10.1109/DISCEX.2003.1194894

[32] Yamanishi K, Takeuchi JI. Discovering outlier filtering rules from unlabeled data: Combining a supervised learner with an unsupervised learner. In: Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Edmonton, Canada; 2001. pp. 389-394. DOI: 10.1145/502512.502570

[33] Yamanishi K, Takeuchi J, Williams G, Milne P. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. Data Mining and Knowledge Discovery. 2004;**8**(3):275-300. DOI: 10.1023/B:DAMI.0000023676.72185.7c

[34] Mahoney MV, Chan PK. Learning nonstationary models of normal network traffic for detecting novel attacks. In: Proceedings of the 8th ACM

SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton, Alberta, Canada: ACM; 23-26 July 2002. pp. 376-386. DOI: 10.1145/775047.775102

[35] Ye N, Emran SM, Chen Q, Vibert S. Multivariate statistical analysis of audit trails for host-based intrusion detection. IEEE Transactions on Computers. 2002;**51**(7):810-820. DOI: 10.1109/TC.2002.1017701

[36] Zhang J, Zulkernine M. Anomaly based network intrusion detection with unsupervised outlier detection. In: Proceedings of the IEEE International Conference on Communications (ICC'06); Istanbul, Turkey; 11-15 June 2006. DOI: 10.1109/ICC.2006.255127

[37] Barbara D, Couto J, Jajodia S, Wu N. ADAM: A testbed for exploring the use of data mining in intrusion detection. In: Proceedings of the ACM SIGMOD; Santa Barbara, CL; 2001. DOI: 10.1145/604264.604268

[38] Zhang J, Zulkernine M. A hybrid network intrusion detection technique using random forests. In: Proceedings of the 1st International Conference on Availability, Reliability, and Security (ARES'06). Vienna, Austria: IEEE; 20-22 April 2006. DOI: 10.1109/ARES.2006.7

[39] Anderson D, Frivold T, Valdes A. Next-generation intrusion detection expert system (NIDES)—A summary. Technical Report SRI-CSL-95-07, SRI; 1995

[40] Agrawal R, Gehrke J, Gunopulos D, Raghavan P. Automatic subspace clustering of high dimensional data for data mining applications. In: Proceedings of ACM SIGMOD. Seattle, WA: ACM; 1998. pp. 94-105. DOI: 10.1145/276305.276314

[41] Sen J, Sengupta I. Autonomous agent-based distributed fault-tolerant

intrusion detection system. In:
Proceedings of the 2nd International
Conference on Distributed Computing
and Internet Technology (ICDCIT'05).
Bhubaneswar, India: Springer; LNCS
Vol. 3186; 22-24 December 2005.
pp. 125-131. DOI: 10.1007/11604655_16

[42] Sen J, Chowdhury PR,
Sengupta I. An intrusion detection
framework in wireless ad hoc
network. In: Proceedings of the
International Conference on Computer
and Communication Engineering
(ICCCE'06); KL, Malaysia; 10-12 May
2006

[43] Sen J, Sengupta I, Chowdhury PR.
An architecture of a distributed
intrusion detection system using
cooperating agents. In: Proceedings
of the International Conference
on Computing and Informatics
(ICOCI'06). KL, Malaysia: IEEE; 6-8
June 2006. pp. 1-6. DOI: 10.1109/
ICOCI.2006.5276474