Volume 2, Issue 4, April 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Novel Method for Intrusion Detection using Data Mining

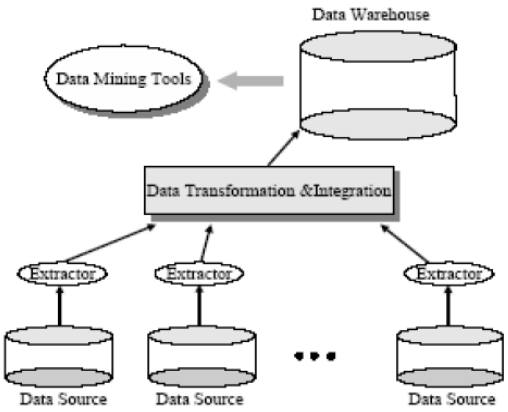
Sherish Johri
IMSEC, Ghaziabad(India)
sherish3@gmail.com

Abstract: ***Intrusion detection is an essential component of the layered computer security mechanisms. It requires accurate and efficient models for analyzing a large amount of system and network audit data. Intrusion detection does not, in general, include prevention of intrusions. In this paper, we focused on data mining techniques to build intrusion detection models. We describe a framework for mining patterns from system and network audit data, and constructing features according to analysis of intrusion patterns. We discuss approaches for improving the run-time efficiency as well as the credibility of detection models. We proposed a data mining approach that we feel can contribute significantly in the attempt to create better and more effective Intrusion Detection Systems.***

Keywords: Data mining, Intrusion detection, Intrusion prevention, Network security, IDS

1. Introduction

The ubiquitous use of computers and computer networks in today's society has made computer network security an international priority. Since it is not technically feasible to build a system with no vulnerabilities, intrusion detection has become an important area of research. Intrusion detection approaches are commonly divided into two categories: misuse detection and anomaly detection [1]. The misuse detection approach attempts to recognize attacks that follow intrusion patterns that have been recognized and reported by experts. Misuse detection systems are vulnerable to intruders who use new patterns of behavior or who mask their illegal behavior to deceive the detection system. Anomaly detection methods were developed to counter this problem. With the anomaly detection approach, one represents patterns of normal behavior, with the assumption that an intrusion can be identified based on some deviation from this normal behavior. When such a deviation is observed, an intrusion alarm is produced.



Data ware housing architecture [22]

Intrusion detection (ID) is a type of security management system for computers and networks. An ID system gathers and analyses information from various areas within a computer or a network to identify possible security breaches.
Intrusion detection functions include:
Monitoring and analysing both user and system activities.
Assessing system and file integrity.
Ability to recognize patterns typical of attacks.
Analysis of abnormal activity patterns
Tracking user policy violations.

ID systems are being developed in response to the increasing number of attacks on major sites and networks. According to webopedia [2] an intrusion detection system (IDS) inspects all inbound and outbound network activity and identifies suspicious patterns that may indicate a network or system attack from someone attempting to break into or compromise a system. More details and information on the various IDS systems and the way they work can be found in [6][13][14][15].

2. Data Mining-What is it??

Data mining (DM), also called Knowledge-Discovery and Data Mining, is the process of automatically searching large volumes of data for patterns using association rules. It is a fairly recent topic in computer science but utilizes many older computational techniques from statistics, information retrieval, machine learning and pattern recognition. Here are a few specific things that data mining might contribute to an intrusion detection project:
Remove normal activity from alarm data to allow analysts to focus on real attacks.

Identify false alarm generators and "bad" sensor signatures.
Find anomalous activity that uncovers a real attack.
Identify long, ongoing patterns (different IP address, same activity).
To accomplish these tasks, data miners use one or more of the following techniques:
Data summarization with statistics, including finding outliers
Visualization: presenting a graphical summary of the data
Clustering of the data into natural categories [Manganaris et al., 2000]
Association rule discovery: defining normal activity and enabling the discovery of anomalies [Clifton and Gengo, 2000; Barbara et al., 2001]
Classification predicting the category to which a particular record belongs.

3. Related work

This section briefly summarizes related experiments on constructing classification models for intrusion detection. These experiments showed the effectiveness of
classification models computed by machine learning programs. Some of the implemented systems that apply data mining techniques in the field of Intrusion Detection are:

ISOA (Information Security Officer's Assistant) [4]: ISOA is a system for monitoring security relevant behavior in computer networks. ISOA serves as the central point for real-time collection and analysis of audit information. When an anomalous situation is identified, associated indicators are triggered. ISOA automates analysis of audit trails, allowing indications and warnings of security threats to be generated in a timely manner so that threats can be countered. ISOA allows a single designated workstation to perform automated security monitoring, analysis and warning.

Distributed Intrusion Detection System (DIDS) [3]: A risk intrusion detection system that aggregates audit reports from a collection of hosts on a single network. Unique to DIDS is its ability to track a user as he establishes connections across the network.

The MINDS System [5]: The Minnesota Intrusion Detection System (MINDS) uses data mining techniques to automatically detect attacks against computer networks and systems. While the long-term objective of MINDS is to address all aspects of intrusion detection, the system currently focuses on two specific issues: – An unsupervised anomaly detection technique that assigns a score to each network connection that reflects how anomalous the connection is, and – An association pattern analysis that summarizes those network connections that are ranked highly anomalous by the anomaly detection module. Experimental results on live network traffic at the University of Minnesota show that the applied anomaly detection techniques are very promising and are successful in automatically detecting several novel intrusions that could not be identified using popular signature-based tools such as SNORT. Furthermore, given the very high volume of connections observed per unit time, association pattern based summarization of novel attacks is quite useful in enabling a security analyst to understand and characterize emerging threats.

Alert Correlation: [8, 9] use correlation techniques to construct "attack scenarios" from low level alerts. [7] also describes a language for modeling alert correlation. [10, 11] describe probabilistic alert correlation. [12] describes use of attack graphs to correlate intrusion event.

4. Architecture for Intrusion Detection:
The security of a computer system is compromised when an intrusion takes place. An intrusion can be defined [16] as ``any set of actions that attempt to compromise the integrity, confidentiality or availability of a resource". Intrusion prevention techniques, such as user authentication (e.g. using passwords or biometrics), avoiding programming errors, and information protection (e.g., encryption) have been used to protect computer systems as a first line of defense. Intrusion prevention alone is not sufficient because as systems become ever more complex, there are always exploitable weakness in the systems due to design and programming errors, or various ``socially engineered" penetration techniques. For example, after it was first reported many years ago, exploitable ``buffer overflow" still exists in some recent system software due to programming errors. The policies that balance convenience versus strict control of a system and information access also make it impossible for an operational system to be completely secure.

Our current architecture for intrusion detection is shown in Figure 1. Network traffic is analyzed by a variety of available sensors. This sensor data is pulled periodically to a central server for conditioning and input to a relational database. HOMER filters events from the sensor data before they are passed on to the classifier and clustering analyses. Data mining tools filter false alarms and identify anomalous behavior in the large amounts of remaining data. A web server is available as a front end to the database if needed, and analysts can launch a number of predefined queries as well as free form SQL queries from this interface. The goal of this operational model is to have all alarms reviewed by human analysts. Without automated support, this task is increasingly difficult due to the volume of alarms. In one recent day at MITRE for example, sensors generated about 3.4 million alarms, of which about 48,000 are labeled priority 1. Attacks and probes can be frequent and noisy, generating thousands of alarms in a day. This can create a burden on the network security analyst, who must perform a triage on the enormous flood of alarms.

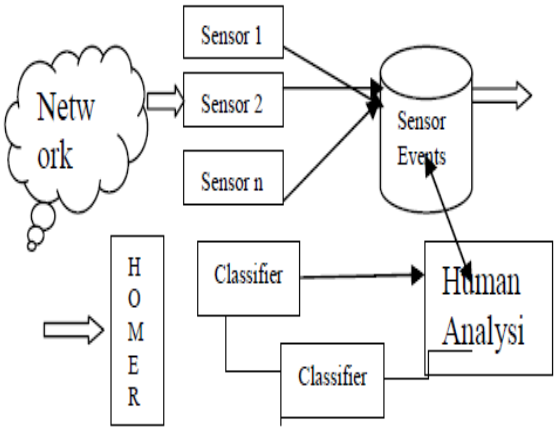


Fig. 1 Overall intrusion detection system

5. Proposed Work:

In this section we propose a data mining technique that could potentially prove to be beneficial to IDSs. The idea is to use biclustering as a tool to analyze network traffic and enhance IDSs. Bi-clustering is the problem of finding a partition of the vectors and a subset of the dimensions such that the projections along those directions of the vectors in each cluster are close to one another. The problem requires the clustering of the vectors and the dimensions imultaneously. The clusters produced by this process are called biclusters. Biclustering measures the similarity across a subset of the experiments, when the testing conditions are heterogeneous. Biclusters may overlap, revealing the role of features in multiple objects and the relations between different objects. The easiest way to approach the problem is by representing the data in a matrix form. Each row represents an object (e.g a traffic trace generated by a process/user) and each column represents a feature (e.g. the Destination Port).Biclustering is now reduced to the problem of finding a subset of the rows and a subset of the columns such that the submatrix induced has the property that each row reads the same string.

	Feature A	Feature B	Feature C	Feature D	Feature E
Process 1	A1	B2	C3	D2	E2
Process 2	A2	B1	C1	D1	E1
Process 3	A1	B2	C1	D2	E2
Process 4	A1	B2	C2	D2	E2
Process 5	A1	B3	C3	D2	E1
Process 6	A2	B1	C2	D3	E1
Process 7	A1	B3	C3	D2	E3

TABLE-1
(BICLUSTERING TECHNIQUE)

Explanation of Biclustering Technique by Example is:

In Table I the rows represents processes (or more accurately the traces they produced) and the columns represent selected features a process trace can have. For simplicity, we consider that each feature has 3 possible discrete values (e.g. feature A can only take values from the set [A1, A2, A3])

By applying Biclustering to the above matrix we find the following 2 clusters:

- {(Process 1, Process 5, Process 7) (A, C, D)}
- {(Process 1, Process 3, Process 4) (A, B, D) }

The first cluster shows that Process 1, 5 and 7 always have the same values for features A, C and D. The second cluster shows that Process 1, 3 and 4 always have the same values for features B, D and E.

If we know in advance that the processes in the matrix are malicious, these process can give us the characteristic feature set for malicious traces. The set can be then used to classify new data collected from the network. Even if know nothing about the processes, this process will not only cluster them, but also show cluster their feature sets. The obtained biclusters could be an effective way to summarize and separate similar processes and analyze them as a group. In general, biclustering can provide valuable knowledge on the relationships between processes and features. More information on biclustering can be found in [17][18][19].

6. Conclusion

Data mining can help improve intrusion detection towards the enhancement of IDS by adding a level of focus to anomaly detection. Our system architecture allows us to support both anomaly detection and misuse detection components at both the individual workstation level and at the network level. We have shown the ways in which data mining has been known to aid the process of Intrusion Detection and the ways in which the various techniques have been applied and evaluated by researchers. Finally, in the last section, we proposed a data mining approach that we feel can contribute significantly in the attempt to create better and more effective Intrusion Detection Systems.

Future work will include expanding the correlation capabilities of our system

Reference:

1. Sundaram, A. 1996. An introduction to intrusion detection. http://www.cs.purdue.edu/cost/archive/data/categ24.html (Accessed 10 November 1999).

2. http://www.webopedia.com.

3. Snapp, S. R., Smaha, S. E., Grance, T., Teal, D. M., ”The DIDS (Distributed Intrusion Detection System) Prototype”, In Proceedings of the USENIX Summer 1992 Technical Conference, pages 227-233, June 1992.

4. Winkler, J. R., Landry, L. C., ”Intrusion and anomaly detection, ISOA update”, In Proceedings of the 15th National Computer Security Conference, pages 272-281, Oct. 1992.

5. ".Data mining for intrusion detection a critical
6. Biswanath Mukherjee,L.Todd Heberlein, Karl N.Levitt, "Network Intrusion Detection",IEEE, June 1994.
7. F. Cuppens and A. Miege, "Alert correlation in a cooperative intrusion detection framework," in Proc. IEEE Symposium on Security and Privacy, May 2002.
8. P. Ning, Y. Cui, and D.S. Reeves, "Constructing attack scenarios through correlation of intrusion alerts," in Proc ACM Computer and Communications Security Conf., 2002.
9. P. Ning and D. Xu, "Learning attack strategies from intrusion alerts," in Proc ACM Computer and Communications Security Conf., 2003.
10. X. Qin and W. Lee, "Statistical causality analysis of INFOSEC alert data," in Proceedings of 6th International Symposium on Recent Advances in Intrusion Detection (RAID 2003), September 2003.
11. X. Qin and W. Lee, "Discovering novel attack strategies from INFOSEC alerts," in Proceedings of the 9th European Symposium on Research in Computer Security (ESORICS 2004), September 2004.
12. S, Noel, E. Robertson, and S. Jajodia, "Correlating intrusion events and building attack scenarios through attack graph distances," in Proceedings of the 20th Annual Computer Security Applications Conference, Tucson, Arizona, December 2004.
13. Presentation on Intrusion Detection Systems, Arian Mavriqi.
review" Klaus Julisch IBM Research
14. Intrusion Detection Methodologies Demystified, Enterasys Networks TM.
15. Protocol Analysis VS Pattern matching in Network and Host IDS, 3rd Generation Intrusion Detection Technology from Network ICE.
16. "The architecture of a network level intrusion detection system. Technical report, R. Heady, G. Luger, A. Maccabe, and M. Servilla. Computer Science Department, University of New Mexico, August 1990.
17. Yang,J., Wang,H., Wang,W., and Yu,P., "Enhanced biclustering on expression data", In Proceedings of the 3rd IEEE Conference on Bioinformatics and Bioengineering (BIBE), pp. 321-327, 2003.
18. Q. Sheng, Y. Moreau, and B. De Moor, "Biclustering microarray data by Gibbs sampling", Bioinformatics,19(Suppl. 2):II196–II205, 2003
19. Mishra N, Ron D., Swaminathan R., "A New Conceptual Clustering Framework", Journal on Machine Learning, Volume 56, Numbers 1-3 / July, 2004, pages 115-151,Springer Netherlands.
20) Kurth thearling Tutorial "An Introduction to Data Mining www.thearling.co
21) Data Mining Approaches for Intrusion Detection" Lee, Wenke and Stolfo, Salvatore.
22) Vuda Sreenivasa Rao, Dr. S Vidyavathi. 'DISTRIBUTED DATA MINING AND MINING MULTI-AGENT DATA'

