Combining Knowledge Discovery and Knowledge Engineering to Build IDSs*

Wenke Lee
wenke@csc.ncsu.edu
Department of Computer Science
North Carolina State University
Raleigh, NC 27695

Salvatore J. Stolfo
sal@cs.columbia.edu
Department of Computer Science
Columbia University
New York, NY 10027

Abstract

We have been developing a data mining (i.e., knowledge discovery) framework, MADAM ID, for Mining Audit Data for Automated Models for Intrusion Detection [LSM98, LSM99b, LSM99a]. The 1998 DARPA Intrusion Detection Evaluation showed that the models produced by MADAM ID performed comparably well with the best purely knowledge-engineered systems. Although our data mining techniques have shown great potentials, it is important recognize the critical roles that domain knowledge, and thus knowledge engineering, play in the process of building ID models. In this paper, we examine why domain knowledge is required in the data mining process, and suggest how to combine knowledge discovery and knowledge engineering to build IDSs.

We first briefly review the main ideas behind MADAM ID. The main components of MADAM ID are classification and meta-classification [CS93] programs, association rules [AIS93] and frequent episodes [MTV95] programs, a feature construction system, and a conversion system that translates off-line learned rules into real-time modules. The end products are concise and intuitive rules that can detect intrusions.

Inductively (and automatically) learned classification rules are more "general" than the manually encoded rules, because classification programs can analyze a large amount of audit data and extract the most general descriptions about an intrusion, whereas a human expert tends to produce very specific rules because he or she can only examine a limited amount of data. As a result, classification rules are often more accurate in detecting not only instances of the "known" intrusions (used in training the rules), but also "new" intrusions that are slight variations of the "known" ones.

A critical requirement for the classification rules to be effective is that an appropriate set of features need to be first constructed and included in the audit records. A focus of our research is thus in automatic "feature construction".

Using MADAM ID, raw audit data is first preprocessed into records with a set of "intrinsic" (i.e., general purposes) features, e.g., duration, source and destination hosts and ports, number of bytes transmitted, etc. Data mining algorithms are then applied to compute the frequent activity patterns,

*This research is supported in part by grants from DARPA (F30602-96-1-0311) and NSF (IRI-96-32225 and CDA-96-25374).

in the forms of association rules and frequent episodes, from the audit records. Association rules describe correlations among system features (e.g., what shell *command* is associated with what *argument*); whereas frequent episodes capture the sequential (temporal) co-occurrences of system events (e.g., what network connections are made within a short time-span). Together, association rules and frequent episodes form the statistical summaries of system activities.

When sufficient normal audit data is gathered, a set of normal frequent patterns can be computed and maintained as baseline. Patterns from audit data of a simulated or real intrusion is then automatically compared with the normal pattern set. The unique "intrusion-only" patterns are then parsed to generate (additional) temporal and statistical features (e.g., for the past 2 seconds, the count of certain types of connections to the same host) into the audit records. Because of these additional features, audit records that contribute to the "intrusion-only" patterns (and hence are likely to be part of the intrusion) are very "different" from the other (normal) audit records. Therefore, with these features, classification rules may be more accurate in detecting the intrusion.

Domain knowledge is required in MADAM ID. Human experts need to first define a basic set of features as the seed for the automatic feature construction process. Our experience in building ID models using *tcpdump* [JLM89] data and *BSM* [Sun95] data also showed that domain knowledge about network and operating systems is required, in the forms of constraints and templates, to direct the data mining programs to efficiently compute only the *useful* patterns and the pattern parsing program to abstract the *anatomy* and *invariant* information about attacks. For extreme "low data volume" attacks, e.g., "land", MADAM ID would fail to compute their "intrusion-only" patterns and thus the appropriate features, due to the statistical nature of the data mining algorithms and the difficulties in setting the correct threshold values. In such cases, human experts need to define the features for the ID models.

Domain knowledge can be systematically incorporated into the feature construction process. First, an extensive set of features can be extracted from existing knowledge-engineered systems such as Bro [Pax98], EMERALD [PN97], IDIOT [KS95], and USTAT [Ilg92]. For example, the "hot" indicators in Bro, the conditions in the EMERALD rules, and the states and conditions in IDIOT and USTAT, etc., can be processed into appropriate forms of feature (i.e., record attribute) definitions for data mining tasks. In addition, the specification the essential attributes (i.e., the unique id) of a network connection or host session can be used to direct data mining programs to compute patterns related to certain connections or sessions. And finally, features of the high confidence rules (i.e., highly accurate rules) can be assigned more *weights* so that they are more likely to be selected by the classification programs into the output rules even if there happens to be some other features that have the same "predictive" power (e.g., *information gain*) on the given training data set. Classification programs tend to generate overly simplified rules and sometimes will not use the most "sensible" features, especially when the training data is not sufficiently representative.

A natural question to ask is "what are the (added) values of data mining systems, such as MADAM ID, if they should utilize existing knowledge engineered systems?" Knowledge-engineered systems are difficult to customize for a new (changed) environment. For example, a rule that tests the frequency of "rejected" connections to the same destination host may need to be fine tuned according to the local network traffic characteristics to select a new frequency threshold. A data mining system can use the gathered local audit data to automatically learn the localized rule given that the relevant features are defined. We believe that the ability for automatic rule adjustments is essential for the rapid and widespread deployment of IDSs. Automatic feature

construction systems, e.g., the one in MADAM ID, can generate features that are aggregates (e.g. functions) on the pre-defined (e.g., knowledge-engineered) features. These additional features are usually more predictive (i.e., they have higher information gain values) and are therefore likely to result in more accurate classification rules. The feature construction process is iterative, and thus very complex (i.e., aggregates of aggregates, etc.) and accurate features can be automatically produced.

We are beginning to study the problem of incorporating "cost" information into intrusion detection models. There are many aspects in the cost models for intrusion detection, for example, the cost (damage) of an intrusion, the (computational) cost of detection (i.e., the cost of computing the features and rules), and the cost of building the ID models (e.g., the cost of the data mining process), etc. A cost matrix can be defined (and measured) by human experts and be imported to a classification program to produce cost-sensitive rules. For example, such matrix can define the (measured) computation costs of the features and the cost-sensitive factors for the intrusion types. A cost-sensitive factor defines the tradeoff between accuracy (predictive power) and cost.

The best way to utilize the off-line learned rules for real-time detection is to automatically covert the rules into modules of existing knowledge-engineered real-time IDSs. MADAM ID employs such a conversion system. In a pilot project with NFR [Net97], we are studying how to automatically generate efficient N-code filters from the learned rules. Again, domain knowledge plays a very important role here. A set of measured low-cost (i.e., efficient) features are used to generate a set of "necessary" conditions, which are 100% confidence association rules with LHSs being the intrusion types and RHSs being the low-cost features. A violation of a necessary condition means that there is no need to compute the high-cost features and check the rules for an intrusion. Therefore, before checking any intrusion detection rule, the list of low-cost necessary conditions can be examined first, and as a result, a lot of rules are "filtered" (no need to be checked). The overall rule checking process is more efficient.

In summary, despite the advantages of using data mining approaches to build intrusion detection models, domain knowledge needs to be properly incorporated. Combining knowledge discovery and knowledge engineering techniques will produce more accurate and efficient intrusion detection models.

References

[AIS93] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, pages 207–216, 1993.

[CS93] P. K. Chan and S. J. Stolfo. Toward parallel and distributed learning by meta-learning. In *AAAI Workshop in Knowledge Discovery in Databases*, pages 227–240, 1993.

[Ilg92] Koral Ilgun. USTAT: A real-time intrusion detection system for Unix. Master's thesis, University of California at Santa Barbara, November 1992.

[JLM89] V. Jacobson, C. Leres, and S. McCanne. *tcpdump*. available via anonymous ftp to ftp.ee.lbl.gov, June 1989.

[KS95] S. Kumar and E. H. Spafford. A software architecture to support misuse intrusion detection. In *Proceedings of the 18th National Information Security Conference*, pages 194–204, 1995.

[LSM98] W. Lee, S. J. Stolfo, and K. W. Mok. Mining audit data to build intrusion detection models. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, New York, NY, August 1998. AAAI Press.

[LSM99a] W. Lee, S. J. Stolfo, and K. W. Mok. A data mining framework for building intrusion detection models. In *Proceedings of the 1999 IEEE Symposium on Security and Privacy*, May 1999.

[LSM99b] W. Lee, S. J. Stolfo, and K. W. Mok. Mining in a data-flow environment: Experience in network intrusion detection. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-99)*, August 1999.

[MTV95] H. Mannila, H. Toivonen, and A. I. Verkamo. Discovering frequent episodes in sequences. In *Proceedings of the 1st International Conference on Knowledge Discovery in Databases and Data Mining*, Montreal, Canada, August 1995.

[Net97] Network Flight Recorder Inc. Network flight recorder. http://www.nfr.com, 1997.

[Pax98] V. Paxson. Bro: A system for detecting network intruders in real-time. In *Proceedings of the 7th USENIX Security Symposium*, San Antonio, TX, 1998.

[PN97] P. A. Porras and P. G. Neumann. EMERALD: Event monitoring enabling responses to anomalous live disturbances. In *National Information Systems Security Conference*, Baltimore MD, October 1997.

[Sun95] SunSoft. *SunSHIELD Basic Security Module Guide*. SunSoft, Mountain View, CA, 1995.

Biography

Wenke Lee received his Ph.D. in Computer Science at Columbia University in 1999. His thesis research involves developing and applying data mining techniques for building intrusion detection models. He has designed and implemented a framework, MADAM ID, for Mining Audit Data for Automated Models for Intrusion Detection. He is currently an assistant professor in Computer Science at North Carolina State University.

Salvatore J. Stolfo is a professor in Computer Science at Columbia University. His research interests include data mining, machine learning, fraud detection, and intrusion detection. He is the PI for the DARPA funded JAM (Java Agents for Meta-learning) project. JAM technologies have been applied to financial fraud detection and intrusion detection. Professor Stolfo will be the program co-chair for the 2000 SIGKDD International Conference on Knowledge Discovery and Data Mining.