Data Mining for Intrusion Detection

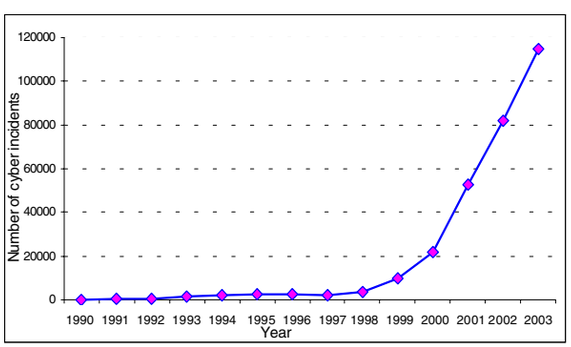
Aleksandar Lazarevic
University of Minnesota, USA

INTRODUCTION

Today computers control power, oil and gas delivery, communication systems, transportation networks, banking and financial services, and various other infrastructure services critical to the functioning of our society. However, as the cost of the information processing and Internet accessibility falls, more and more organizations are becoming vulnerable to a wide variety of cyber threats. According to a recent survey by CERT/CC (Computer Emergency Response Team/Coordination Center), the rate of cyber attacks has been more than doubling every year in recent times (*Figure 1*). In addition, the severity and sophistication of the attacks are also growing. For example, Slammer/Sapphire Worm was the fastest computer worm in history. As it began spreading throughout the Internet, it doubled in size every 8.5 seconds and infected at least 75,000 hosts causing network outages and unforeseen consequences such as canceled airline flights, interference with elections, and ATM failures (Moore, 2003).

It has become increasingly important to make our information systems, especially those used for critical functions in the military and commercial sectors, resistant to and tolerant of such attacks. The conventional approach for securing computer systems is to design security mechanisms, such as firewalls, authentication mechanisms, and Virtual Private Networks (VPN) that create a protective "shield" around them. However, such security mechanisms almost always have inevitable vulnerabilities and they are usually not sufficient to ensure complete security of the infrastructure and to ward off attacks that are continually being adapted to exploit the system's weaknesses often caused by careless design and implementation flaws. This has created the need for security technology that can monitor systems and identify computer attacks. This component is called intrusion detection and is a complementary to conventional security mechanisms. This article provides an overview of current status of research in intrusion detection based on data mining.

Figure 1. Growth rate of cyber incidents reported to Computer Emergency Response Team/Coordination Center (CERT/CC)



BACKGROUND

Intrusion detection includes identifying a set of malicious actions that compromise the integrity, confidentiality, and availability of information resources. An Intrusion Detection System (IDS) can be defined as a combination of software and/or hardware components that monitors computer systems and raises an alarm when an intrusion happens.

Traditional intrusion detection systems are based on extensive knowledge of signatures of known attacks. However, the signature database has to be manually revised for each new type of intrusion that is discovered. In addition, signature-based methods cannot detect emerging cyber threats, since by their very nature these threats are launched using previously unknown attacks. Finally, very often there is substantial latency in deployment of newly created signatures. All these limitations have led to an increasing interest in intrusion detection techniques based upon data mining.

The tremendous increase of novel cyber attacks has made data mining based intrusion detection techniques extremely useful in their detection. Data mining techniques for intrusion detection generally fall into one of three categories; misuse detection, anomaly detection and summarization of monitored data.

MAIN THRUST

Before applying data mining techniques to the problem of intrusion detection, the data has to be collected. Different types of data can be collected about informa-

tion systems (e.g., tcpdump and netflow data for network intrusion detection, syslogs or system calls for host intrusion detection). However, such collected data is often available in a raw format and needs to be processed in order to be used in data mining techniques. For example, in MADAM ID project (Lee, 2000, 2001) at Columbia University, association rules and frequent episodes were extracted from network connection records to construct three groups of features: (i) content-based features that describe intrinsic characteristics of a network connection (e.g., number of packets, acknowledgments, data bytes from source to destination), (ii) time-based traffic features that compute the number of connections in some recent time interval (e.g., last few seconds) and (iii) connection based features that compute the number of connections from a specific source to a specific destination in the last N connections (e.g., N = 1000).

When the feature construction step is complete, obtained features may be used in any data mining technique.

Misuse Detection

In misuse detection based on data mining, each instance in a data set is labeled as "normal" or "attack/intrusion" and a learning algorithm is trained over the labeled data. These techniques are able to automatically retrain intrusion detection models on different input data that include new types of attacks, as long as they have been labeled appropriately. Unlike signature-based intrusion detection systems, data mining based misuse detection models are created automatically, and can be more sophisticated and precise than manually created signatures. In spite of the fact that misuse detection models have high degree of accuracy in detecting known attacks and their variations, their obvious drawback is the inability to detect attacks whose instances have not yet been observed. In addition, labeling data instances as normal or intrusive may require enormous time for many human experts.

Since standard data mining techniques are not directly applicable to the problem of intrusion detection due to dealing with skewed class distribution (attacks/intrusions correspond to a class of interest that is much smaller, i.e., rarer, than the class representing normal behavior) and learning from data streams (attacks/intrusions very often represent sequence of events), a number of researchers have developed specially designed data mining algorithms that are suitable for intrusion detection. Research in misuse detection has focused mainly on classification of network intrusions using various standard data mining algorithms (Barbara, 2001; Ghosh, 1999; Lee, 2001; Sinclair, 1999), rare class predictive models (Joshi, 2001) and association rules (Barbara, 2001; Lee, 2000; Manganaris, 2000).

MADAM ID (Lee, 2000, 2001) was one of the first projects that applied data mining techniques to the intrusion detection problem. In addition to the standard features that were available directly from the network traffic (e.g., duration, start time, service), three groups of constructed features were also used by the RIPPER algorithm to learn intrusion detection rules from DARPA 1998 data set (Lippmann, 1999). Other classification algorithms that are applied to the intrusion detection problem include standard decision trees (Bloedorn, 2001; Sinclair, 1999), modified nearest neighbor algorithms (Ye, 2001b), fuzzy association rules (Bridges, 2000), neural networks (Dao, 2002; Lippman, 2000a), naïve Bayes classifiers (Schultz, 2001), genetic algorithms (Bridges, 2000), genetic programming (Mukkamala, 2003a), and etcetera. Most of these approaches attempt to directly apply specified standard techniques to publicly available intrusion detection data sets (Lippmann, 1999, 2000b), assuming that the labels for normal and intrusive behavior are already known. Since this is not realistic assumption, misuse detection based on data mining has not been very successful in practice.

Anomaly Detection

Anomaly detection creates profiles of normal "legitimate" computer activity (e.g., normal behavior of users, hosts, or network connections) using different techniques and then uses a variety of measures to detect deviations from defined normal behavior as potential anomaly. Anomaly detection models often learn from a set of "normal" (attack-free) data, but this also requires cleaning data from attacks and labeling only normal data records. Nevertheless, other anomaly detection techniques detect anomalous behavior without using any knowledge about the training data. Such models typically assume that the data records that do not belong to the majority behavior correspond to anomalies.

The major benefit of anomaly detection algorithms is their ability to potentially recognize unforeseen and emerging cyber attacks. However, their major limitation is potentially high false alarm rate, since deviations detected by anomaly detection algorithms may not necessarily represent actual attacks, but new or unusual, but still legitimate, network behavior.

Anomaly detection algorithms can be classified into several groups: (i) statistical methods; (ii) rule-based methods; (iii) distance-based methods; (iv) profiling methods; and (v) model-based approaches (Lazarevic, 2004). Although anomaly detection algorithms are quite diverse in nature, and thus may fit into more than one proposed category, most of them employ certain data mining or artificial intelligence techniques.