Data Mining in Security Applications

Aleksandar Lazarevic
United Technologies Research Center, USA

INTRODUCTION

In recent years, research in many security areas has gained a lot of interest among scientists in academia, industry, military and governmental organizations. Researchers have been investigating many advanced technologies to effectively solve acute security problems. Data mining has certainly been one of the most explored technologies successfully applied in many security applications ranging from computer and physical security and intrusion detection to cyber terrorism and homeland security. For example, in the context of homeland security, data mining can be a potential means to identify terrorist activities, such as money transfers and communications, and to identify and track individual terrorists themselves, such as through travel and immigration records (Seifert, 2007). In another data mining's success story related to security, credit card fraud detection, all major credit card companies mine their transaction databases, looking for spending patterns that indicate a stolen card. In addition, data mining has also effectively been utilized in many physical security systems (e.g. in efficient system design tools, sensor fusion for false alarm reduction) and video surveillance applications, where many data mining based algorithms have been proposed to detect motion or intruder at monitored sites or to detect suspicious trajectories at public places.

This chapter provides an overview of current status of data mining based research in several security applications including cyber security and intrusion detection, physical security and video surveillance.

BACKGROUND

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. It has become increasingly important recently 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 security mechanisms, such as firewalls, authentication mechanisms, Virtual Private Networks (VPN) almost always have inevitable vulnerabilities and they are usually insufficient to ensure complete security of the infrastructure and to ward off attacks that are continually being adapted to exploit the system's weaknesses. This has created the need for security technology, called intrusion detection that includes identifying malicious actions that compromise the integrity, confidentiality, and availability of information resources.

MAIN THRUST OF THE CHAPTER

Cyber Security and Intrusion Detection

Traditional intrusion detection systems (IDSs) are based on extensive knowledge of signatures (rule descriptions) 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. These limitations have led to an increasing interest in intrusion detection techniques based upon data mining. Data mining techniques for cyber security and intrusion detection generally fall into one of two categories: misuse detection, and anomaly detection.

Misuse Detection

In misuse detection techniques, each instance in a data set is labeled as 'normal' or 'attack/intrusion' and a learning algorithm is trained over the labeled data. Unlike signature-based IDSs, 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 skewed class distribution (attacks/intrusions correspond to a much smaller, i.e. rarer class, than the class representing normal behavior) and streaming nature of data (attacks/intrusions very often represent sequence of events), a number of researchers have developed specially designed data mining algorithms suitable for intrusion detection. Research in misuse detection has focused mainly on classifying network intrusions using various standard data mining algorithms (Barbara, 2001; Lee, 2001), rare class predictive models (Joshi, 2001) and association rules (Barbara, 2001; Lee, 2000; Manganaris, 2000).

MADAM ID (Lee, 2000; Lee, 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 (content-based features that describe intrinsic characteristics of a network connection (e.g. number of packets, acknowledgments, data bytes from source to destination), time-based traffic features that compute the number of connections in some recent time interval and connection based features that compute the number of connections from a specific source to a specific destination in the last N connections) were also used by the RIPPER algorithm to learn intrusion detection rules. Other classification algorithms that are applied to the intrusion detection problem include standard decision trees (Bloedorn, 2001), modified nearest neighbor algorithms (Ye, 2001b), fuzzy association rules (Bridges, 2000), neural networks (Dao, 2002; Kumar, 2007; Zhang, 2005), support vector machines (Chen, 2005; Kim, 2005), naïve Bayes classifiers (Bosin, 2005; Schultz, 2001), genetic algorithms (Bridges, 2000; Kim, 2005; Li, 2004), genetic programming (Mukkamala, 2003), etc. Most of these approaches attempt to directly apply specified standard techniques to publicly available intrusion detection data sets (Lippmann, 1999; Lippmann, 2000), 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 (regular e-mail reading, web browsing, using specific software), 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 detected deviations 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, 2005b). Although anomaly detection algorithms are quite diverse in nature, and thus may fit into more than one proposed category, most of them employ certain artificial intelligence techniques.

Statistical methods. Statistical methods monitor the user or system behavior by measuring certain variables over time (e.g. login and logout time of each session). The basic models keep averages of these variables and detect whether thresholds are exceeded based on the standard deviation of the variable. More advanced statistical models compute profiles of long-term and short-term user activities by employing different techniques, such as Chi-square (χ^2) statistics (Ye, 2001a), probabilistic modeling (Yamanishi, 2000), Markov-chain models (Zanero, 2006) and likelihood of data distributions (Eskin, 2000).

Distance based methods. Most statistical approaches have limitation when detecting outliers in higher dimensional spaces, since it becomes increasingly difficult and inaccurate to estimate the multidimensional distributions of the data points. Distance based approaches attempt to overcome these limitations by