Chapter XLIV
The Role of Data Mining in Intrusion Detection Technology

Amalia Agathou
University of the Aegean, Greece

Theodoros Tzouramanis
University of the Aegean, Greece

INTRODUCTION

Over the past few years, the Internet has changed computing as we know it. The more possibilities and opportunities develop, the more systems are subject to attack by intruders. Thus, the big question is about how to recognize and handle subversion attempts. One answer is to undertake the prevention of subversion itself by building a completely secure system. However, the complete prevention of breaches of security does not yet appear to be possible to achieve. Therefore these intrusion attempts need to be detected as soon as possible (preferably in real time) so that appropriate action might be taken to repair the damage. This is what an intrusion detection system (IDS) does.

IDSs monitor and analyze the events occurring in a computer system in order to detect signs of security problems. However, intrusion detection technology has not yet reached perfection. This fact has provided data mining with the opportunity to make several important contributions and improvements to the field of IDS technology (Julisch, 2002).

BACKGROUND

IDSs are systems that aim at the detection of subversion and the prevention of similar attacks in the future (Sundaram, 1996). Therefore, an IDS identifies evidence of intrusions, either while they are in progress or after the fact. The most popular way to detect intrusions has been the use of the audit data generated by the operating system. An audit trail is a record of activities on a system that are logged to a file in chronologically sorted order. These data may be collected in many ways, but their sources are usually network activity and/or host-based logs.

Since almost all activities are logged on a system, a manual inspection of these logs may detect intrusions. However, too much data are collected, making manual analysis impossible to be usefully analyzed for intrusions. What IDSs achieve is the automation of the data analysis process. It is thus possible to establish the guilt of the intruders and to detect unauthorized and subversive user activity.

Postmortem analysis of the audit data is significant as it helps to determine the extent of the damage and to identify intruders so that steps may be taken to overcome system weakness.

Kemmer and Vigna (2002) describe an IDS as composed of several components: sensors for the capture of events and for their storage as audit data; an engine for the production of alarm signals upon detection of a potential intrusion, detected from the processing of the audit data captured; and a site security officer (SSO) for the reception of the alarms and for appropriate response.

Data Collection Issues

Achieving reliable and complete data collection about the target system's activities is a complex issue. As Kemmer and Vigna (2002) state, most operating systems offer some form of auditing that provides an operations log for different users.

These data might be collected at the lowest possible level, resulting in the collection of rather large quantities of system activity information to be analyzed for intrusions. To alleviate this problem, the use of random sampling has been suggested; however, this could mean that certain types of attacks may stay undetected. The problem is further complicated by the need to allow for differences in the data as a result of special circumstances such as holidays and other factors.

Finally, depending on a specific IDS solution and its correlation engine, a storage period for current audit files should be appropriately set. For retrieval analysis purposes, archive files should be stored as copies.

Detection Techniques

Traditionally, there are two basic categories of intrusion detection techniques: anomaly detection and misuse detection. Most current intrusion detection systems use one or both of these two approaches.

According to Lee and Stolfo (1998), the anomaly detection model devises a set of statistical metrics that model the behavior of an entity, usually a user, a group of users, or a host computer, interpreting deviations from this behavior profile as a problem. The profile of a user entity, for instance, may include information such as the CPU (central processing unit) usage and the frequency of system commands during a user log-in session. The IDS monitors the operation of a computer system and constantly compares the profile of, say, a current user session with the one stored in its database. If it detects a high deviation from the normal behavior, it signals an alarm to the system security officer. Deviation from a profile can be computed as the weighted sum of the deviations of the constituent statistical measures. Stored profiles are constantly being updated so that shifts of normal behavior are accounted for.

The misuse detection model, on the other hand, uses specifically known patterns of unauthorized behavior to predict and detect subsequent similar attempts. They contain attack descriptions (or signatures; Lee et al., 2000) and match them against the audit data stream, looking for evidence of known attacks. Depending on the robustness and seriousness of a signature that is triggered, some form of alarm, either a response or notification, should be sent to the proper authorities. An obvious difficulty in this architecture is the need for the constant updating of the rule base as new attack methods become known (Julisch, 2002).

Public Domain for Intrusion Detection

The fast development of information and communication technology over the past few years