61

Data Mining for Intrusion Detection

Anoop Singhal[1] and Sushil Jajodia[2]

[1] Center for Secure Information Systems,
George Mason University, Fairfax, VA 22030-4444
[2] Center for Secure Information Systems,
George Mason University, Fairfax, VA 22030-4444

Summary. Data Mining Techniques have been successfully applied in many different fields including marketing, manufacturing, fraud detection and network management. Over the past years there is a lot of interest in security technologies such as intrusion detection, cryptography, authentication and firewalls. This chapter discusses the application of Data Mining techniques to computer security. Conclusions are drawn and directions for future research are suggested.

Key words: computer security, intrusion detection, data warehouse, alert correlation

61.1 Introduction

Computer security is of importance to a wide variety of practical domains ranging from banking industry to multinational corporations, from space exploration to the intelligence community and so on. The following principles are generally accepted as the foundation of a good security solution:

- Authentication: The process of establishing the validity of a claimed identity.
- Authorization: The process of determining whether a validated entity is allowed access to a resource based on attributes, predicates, or context.
- Integrity: The prevention of modification or destruction of an asset by an unauthorized user.
- Availability: The protection of assets from denial-of-service threats that might impact system availability.
- Confidentiality: The property of non-disclosure of information to unauthorized users.
- Auditing: The property of logging all system activities at a sufficient level so that events can be reconstructed if it is required.

Intrusion detection is the process of monitoring and analyzing the events occurring in a computer system in order to detect signs of security problems. Over the past several years,

O. Maimon, L. Rokach (eds.), *Data Mining and Knowledge Discovery Handbook*, 2nd ed.,
DOI 10.1007/978-0-387-09823-4_61, © Springer Science+Business Media, LLC 2010

intrusion detection and other security technologies such as cryptography, authentication and firewalls have increasingly gained in importance. There is a lot of interest in applying Data Mining techniques to intrusion detection. This chapter gives a critical summary of Data Mining research for intrusion detection. We first give the basics of Data Mining techniques. We then survey a list of research projects that apply Data Mining techniques to intrusion detection. We then suggest new directions for research and then give our conclusions.

61.2 Data Mining Basics

Recent progress in scientific and engineering applications has accumulated huge volumes of data. The fast growing, tremendous amount of data, collected and stored in large databases has far exceeded our human ability to comprehend it without proper tools. It is estimated that the total database size for a retail store chain such as Walmart will exceed 1 Petabyte (1K Terabyte) by 2005. Similarly, the scope, coverage and volume of digital geographic data sets and multidimensional data have grown rapidly in recent years. These data sets include digital data of all sorts created and disseminated by government and private agencies on land use, climate data and vast amounts of data acquired through remote sensing systems and other monitoring devices. It is estimated that multimedia data is growing at about 70% per year. Therefore, there is a critical need of data analysis systems that can automatically analyze the data, to summarize it and predict future trends. Data Mining is a necessary technology for collecting information from distributed databases and then performing data analysis.

The process of knowledge discovery in databases is explained in Figure 61.1 and it consists of the following steps (Han and Kamber, 2000):

1. Data cleaning to remove noise and inconsistencies.
2. Data integration to get data from multiple sources.
3. Data selection step where data relevant for the task is retrieved.
4. Data transformation step where data is transformed into an appropriate form for data analysis.
5. Data Analysis where complex queries are executed for in depth analysis.

The following are different kinds of techniques and algorithms that data mining can provide:

Association Analysis: This involves discovery of *association rules* showing attribute-value conditions that occur frequently together in a given set of data. This is used frequently for market basket or transaction data analysis. For example, the following rule says that if a customer is in age group 20 to 29 years and income is greater than 40K/year then he or she is likely to buy a DVD player.

Age(X, "20-29") & income(X, ">40K") => buys (X, "DVD player")

[support = 2% , confidence = 60%]

Rule *support* and *confidence* are two measures of rule interestingness. A support of 2% means that 2% of all transactions under analysis show that this rule is true. A confidence of 60% means that among all customers in the age group 20-29 and income greater than 40K, 60% of them bought DVD players.

A popular algorithm for discovering association rules is the **Apriori** method. This algorithm uses an iterative approach known as *level-wise* search where k-itemsets are used to explore (k+1) itemsets. Association rules are widely used for prediction.

Classification and Prediction: Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future

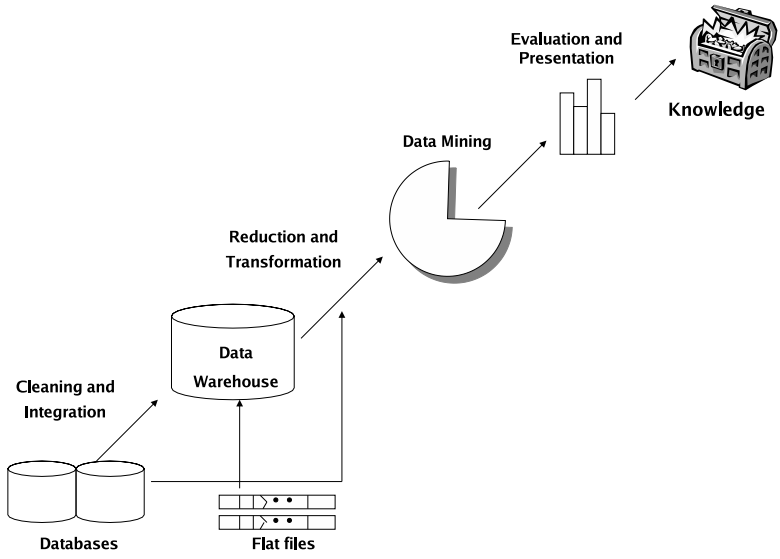

Fig. 61.1. The process of Knowledge Discovery

data trends. For example, a classification model can be built to categorize bank loan applications as either safe or risky. A prediction model can be built to predict the expenditures of potential customers on computer equipment given their income and occupation. Some of the basic techniques for data classification are decision tree induction, Bayesian classification and neural networks.

These techniques find a set of models that describe the different *classes* of objects. These models can be used to predict the class of an object for which the class is unknown. The derived model can be represented as rules (IF-THEN), decision trees or other formulae.

Clustering: This involves grouping objects so that objects within a cluster have high similarity but are very dissimilar to objects in other clusters. Clustering is based on the principle of *maximizing the intraclass similarity and minimizing the interclass similarity.*

In business, clustering can be used to identify customer groups based on their purchasing patterns. It can also be used to help classify documents on the web for information discovery. Due to the large amount of data collected, cluster analysis has recently become a highly active topic in Data Mining research. As a branch of statistics, cluster analysis has been extensively studied for many years, focusing primarily on *distance based cluster analysis.* These techniques have been built into statistical analysis packages such as S-PLUS and SAS. In machine learning, clustering is an example of *unsupervised learning*. For this reason clustering is an example of *learning by observation.*

Outlier Analysis: A database may contain data objects that do not comply with the general model or behavior of data. These data objects are called *outliers.* Most Data Mining methods discard outliers as noise or exceptions. These outliers are useful for applications such as fraud detection and network intrusion detection. The analysis of outlier data is referred to as *outlier mining.* We will describe some intrusion detection systems that use outlier analysis.

Outliers may be detected using statistical tests that assume a distribution or probability model for the data, or using distance measures where objects that are a substantial distance from other clusters are consider outliers.

61.3 Data Mining Meets Intrusion Detection

Since the cost of information processing and Internet accessibility is dropping, more and more organizations are becoming vulnerable to a wide variety of cyber threats. According to a recent survey by CERT, the rate of cyber attacks has been doubling every year in recent times. Therefore, it has become increasingly important to make our information systems, especially those used for critical functions such as military and commercial purpose, resistant to and tolerant of such attacks. Intrusion Detection Systems (IDS) are an integral part of any security package of a modern networked information system. An IDS detects intrusions by monitoring a network or system and analyzing an audit stream collected from the network or system to look for clues of malicious behavior.

Intrusion detection systems can be classified into the following two categories:

- Misuse Detection: This method finds intrusions by monitoring network traffic in search of direct matches to known patterns of attack (called signatures or rules). A disadvantage of this approach is that it can only detect intrusions that match a pre-defined rule. One advantage of these systems is that they have low false alarm rates.
- Anomaly Detection: In this approach, the system defines the expected behavior of the network in advance. The profile of normal behavior is built using techniques that include statistical methods, association rules and neural networks. Any significant deviations from this expected behavior are reported as possible attacks. In principle, the primary advantage of anomaly based detection is the ability to detect novel attacks for which signatures have not been defined yet. However, in practice, this is difficult to achieve because it is hard to obtain accurate and comprehensive profiles of normal behavior. This makes an anomaly detection system generate too many false alarms and it can be very time consuming and labor intensive to sift through this data.

Intrusion Detection Systems (IDS) can also be categorized according to the kind of information they analyze. This leads to the distinction between *host-based* and *network-based* IDSs. A host based IDS analyzes host-bound audit sources such as operating system audit trails, system logs or application logs. Since host based systems directly monitor the host data files and operating system processes, they can determine exactly which host resources are targets of a particular attack. Due to the rapid development of computer networks, traditional single host intrusion detection systems have been modified to monitor a number of hosts on a network. They transfer the monitored information from multiple monitored hosts to a central site for processing. These are termed as distributed intrusion detection systems.

A network based IDS analyzes network packets that are captured on a network. This involves placing a set of traffic sensors within the network. The sensors typically perform local analysis and detection and report suspicious events to a central location.

Recently, there is a great interest in application of Data Mining techniques to intrusion detection systems. The problem of intrusion detection can be reduced to a Data Mining task of classifying data. Briefly, one is given a set of data points belonging to different classes (normal activity, different attacks) and aims to separate them as accurately as possible by means of a model. This section gives a summary of the current research project in this area.

61.3.1 ADAM

The ADAM project at George Mason University (Barbara *et al*., 2001, Barbara*et al*., 2001) is a network-based anomaly detection system. ADAM learns normal network behavior from attack-free training data and represents it as a set of association rules, the so called profile. At run time, the connection records of past delta seconds are continuously mined for new association rules that are not contained in the profile.

ADAM is an anomaly detection system. It is composed of three modules: a preprocessing engine, a mining engine and a classification engine. The preprocessing engine sniffs TCP/IP traffic data and extracts information from the header of each connection according to a predefined schema. The mining engine applies mining association rules to the connection records. It works in two modes: training mode and detecting mode. In training mode, the mining engine builds a profile of the users and systems normal behavior and generates association rules that are used to train the classification engine. In detecting mode, the mining engine mines unexpected association rules that are different from the profile. The classification engine will classify the unexpected association rules into normal and abnormal events. Some abnormal events can be further classified as attacks. Although mining of association rules has used previously to detect intrusions in audit trail data, the ADAM system is unique in the following ways:

- It is on-line; it uses an incremental mining (on-line mining) which does not look at a batch of TCP connections, but rather uses a sliding window of time to find the suspicious rules within that window.
- It is an anomaly detection system that aims to categorize using Data Mining the rules that govern misuse of a system. For this, the technique builds, apriori, a profile of "normal" rules, obtained by mining past periods of time in which there were no attacks. Any rule discovered during the on-line mining that also belongs to this profile is ignored, assuming that it corresponds to a normal behavior.

Figures 61.2 and 61.3 show the basic architecture of ADAM. ADAM performs its task in two phases. In the training phase, ADAM uses a data stream for which it knows where the attacks are located. The attack free parts of the stream are fed into a module that performs off-line association rules discovery. The output of this module is a profile of rules that we call "normal" i.e. it provides the behavior during periods when there are no attacks. The profile along with the training data set is also fed into a module that uses a combination of dynamic, on line algorithm for association rules, whose output consists of frequent item sets that characterize attacks to the system. These item sets are used as a classifier or decision tree. This whole phase takes place off-line before we use the system to detect attacks.

The second phase of ADAM in which we actually detect attacks is shown in the figure below. Again, the on-line association rules mining algorithm is used to process a window of current connections. Suspicious connections are flagged and sent along with their feature vectors to the trained classifier, where they are labeled as attacks, false alarms or unknown. When, the classifier labels connections as false alarms, it is filtering them out of the attacks set and avoiding passing these alerts to the security officer. The last class, i.e. unknown is reserved for the events whose exact nature cannot be confirmed by the classifier. These events are also considered as attacks and they are included in the set of alerts that are passed to the security officer.

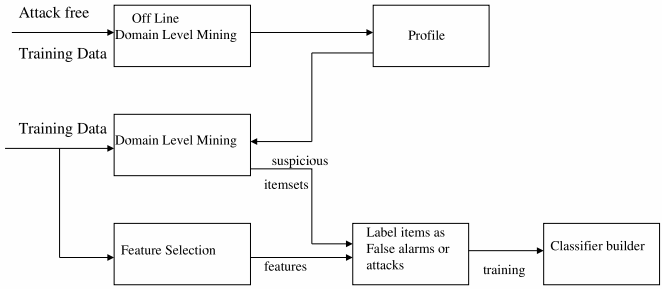

Fig. 61.2. The Training Phase of ADAM

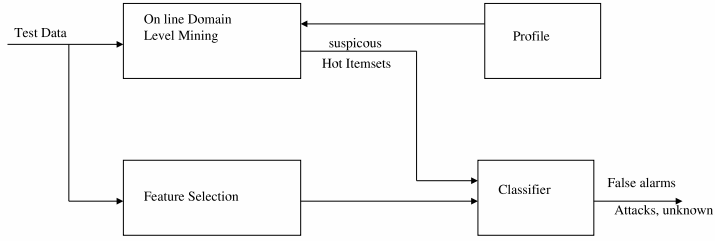

Fig. 61.3. The Intrusion Detection Phase of ADAM

61.3.2 MADAM ID

The MADAM ID project at Columbia University (Lee, 1998, Lee *et al.*, 1998) has shown how Data Mining techniques can be used to construct an IDS in a more systematic and automated manner. Specifically, the approach used by MADAM ID is to learn classifiers that distinguish between intrusions and normal activities. Unfortunately, classifiers can perform really poorly when they have to rely on attributes that are not predictive of the target concept. Therefore, MADAM ID proposes association rules and frequent episode rules as means to construct additional more predictive attributes. These attributes are termed as *features.*

We will describe briefly how MADAM ID is used to construct network based misuse detection systems. First all network traffic is preprocessed to create *connection records.* The attributes of connection records are intrinsic connection characteristics such as source host, the destination host, the source and destination posts, the start time, the duration, header flags and so on. In the case of TCP/IP networks, connection records summarize TCP sessions.

The most important characteristic of MADAM ID is that it *learns* a misuse detection model from examples. In order to use MADAM ID, one needs a large set of connection records that have already been classified into "normal records" or some kind of attacks. MADAM ID proceeds in two steps. In the first step it does *feature construction* in which some additional features are constructed that are considered useful for doing the analysis. One example for this step is to calculate the count of the number of connections that have been initiated during the last two seconds to the same destination host as the current host. The feature construction step is followed by the *classifier learning* step. It consists of the following process:

1. The training connection records are partitioned into two sets, namely *normal connection records* and *intrusion connection records.*
2. Association rules and frequent episode rules are mined separately from the normal connection records and from the intrusion connection records. The resulting patterns are compared and all patterns that are exclusively contained in the intrusion connection records are collected to form *the intrusion only* patterns.
3. The intrusion only patterns are used to derive additional attributes such as count or percentage of connection records that share some attribute values with the current connection records.
4. A classifier is learned that distinguishes normal connection records from intrusion connection records, This classifier is the end product of MADAM ID.

61.3.3 MINDS

The MINDS project (Ertoz *et al.*, 2003, Kumar *et al.*, 2003) at University of Minnesota uses a suite of Data Mining techniques to automatically detect attacks against computer networks and systems. Their system uses an anomaly detection technique to assign a score to each connection to determine how anomalous the connection is compared to normal network traffic. Their experiments have shown that anomaly detection algorithms can be successful in detecting numerous novel intrusions that could not be identified using widely popular tools such as SNORT.

Input to MINDS is Netflow data that is collected using Netflow tools. The netflow data contains packet header information i.e. they do not capture message contents. Netflow data for each 10 minute window which typically results in 1 to 2 million records is stored in a flat file. The analyst uses MINDS to analyze these 10 minute data files in a batch mode. The first step in MINDS involves constructing features that are used in the Data Mining analysis.

Basic features include source IP address and port, destination IP address and port, protocol, flags, number of bytes and number of packets. Derived features include time-window and connection window based features. After the feature construction step, the data is fed into the MINDS anomaly detection module that uses an outlier detection algorithm to assign an anomaly score to each network connection. A human analyst then has to look at only the most anomalous connections to determine if they are actual attacks or other interesting behavior.

MINDS uses a density based outlier detection scheme for anomaly detection. The reader is referred to (Ertoz *et al*., 2003) for a more detailed overview of their research. MINDS assigns a degree of being an outlier to each data point which is called the local outlier factor (LOF). The output of the anomaly detector contains the original Netflow data with the addition of the anomaly score and relative contribution of the different attributes to that score. The analyst typically looks at only the top few connections that have the highest anomaly scores. The researchers of MINDS have their system to analyze the University of Minnesota network traffic. They have been successful in detecting scanning activities, worms and non standard behavior such as policy violations and insider attacks.

61.3.4 Clustering of Unlabeled ID

Traditional anomaly detection systems require "clean" training data in order to learn the model of normal behavior. A major drawback of these systems is that clean training data is not easily available. To overcome this weakness, recent research has investigated the possibility of training anomaly detection systems over noisy data (Portnoy *et al*., 2001). Anomaly detection over noisy data makes two key assumptions about the training data. First, the number of normal elements in the training data is assumed to be significantly larger than the number of anomalous elements. Secondly, anomalous elements are assumed to be qualitatively different from normal ones. Then, given that anomalies are both rare and different, they are expected to appear as outliers that stand out from the normal baseline data. Portnoy et al. (Portnoy *et al*., 2001) apply clustering to the training data. Here the hope is that intrusive elements will bundle with other intrusive elements whereas normal elements will bundle with other normal ones. Moreover, as intrusive elements are assumed to be rare, they should end up in small clusters. Thus, all small clusters are assumed to contain intrusions/anomalies, whereas large clusters are assumed to represent normal activities. At run time, new elements are compared against all clusters and the most similar cluster determines the new element's classification as either "normal" or "intrusive".

61.3.5 Alert Correlation

Correlation techniques from multiple sensors for large networks is described in (Ning *et al*., 2002,Ning and Xu, 2003). A language for modeling alert correlation is described in (Cuppens and Miege, 2002). Traditional IDS systems focus on low level alerts and they raise alerts independently though there may be a logical connection between them. In case of attacks, the number of alerts that are generated become unmanageable. As a result, it is difficult for human users to understand the alerts and take appropriate actions. Ning et al. present a practical method for constructing attack scenarios through alert correlation, using prerequisites and consequences of intrusions. Their approach is based on the observation that in a series of attacks, alerts are not isolated, but related as different stages, with earlier stages preparing for the later ones. They proposed a formal framework to represent alerts with their prerequisites and consequences using the concept of *hyper-alerts.* They evaluated their approach using the 2000 DARPA intrusion detection scenario specific datasets.

61.4 Conclusions and Future Research Directions

In this chapter, we reviewed the application of Data Mining techniques to the area of computer security. Data Mining is primarily being used to detect intrusions rather than to discover new knowledge about the nature of attacks. Moreover, most research is based on strong assumptions that complicate building of practical applications. First, it is assumed that labeled training data is readily available, and second it is assumed that this data is of high quality. Different authors have remarked that in many cases, it is not easy to obtain labeled data. Even if one could obtain labeled training data by simulating intrusions, there are many problems with this approach. Additionally, attack simulation limits the approach to the set of known attacks. We think that the difficulties associated with the generation of high quality training data will make it difficult to apply Data Mining techniques that depend on availability of high quality labeled training data. Finally, Data Mining in intrusion detection focuses on a small subset of possible applications. Interesting future applications of Data Mining might include the discovery of new attacks, the development of better IDS signatures and the construction of alarm correlation systems.

For future research, it should be possible to focus more on the KDD process and detection of novel attacks. It is known that attackers use a similar strategy to attack in the future as what they used in the past. The current IDSs can only detect a fraction of these attacks. There are new attacks that are hidden in the audit logs, and it would be useful to see how Data Mining can be used to detect these attacks.

Data Mining can also be applied to improve IDS signatures. IDS vendors can run their systems in operational environment where all alarms and audit logs are collected. Then, Data Mining can be used to search for audit log patterns that are closely related with particular alarms. This might lead to new knowledge as to why false positives arise and how they can be avoided.

Finally, Data Mining projects should focus on the construction of alarm correlation systems. Traditional intrusion detection systems focus on low level alerts and they raise alerts independently even though there is a logical connection among them. More work needs to be done on alert correlation techniques that can construct "attack strategies" and facilitate intrusion analysis. One way is to store data from multiple sources in a data warehouse and then perform data analysis. Alert correlation techniques will have several advantages. First, it will provide a high level representation of the alerts along with a temporal relationship of the sequence in which these alerts occurred. Second, it will provide a way to distinguish a true alert from a false alert. We think that true alerts are likely to be correlated with other alerts whereas false alerts will tend to be random and, therefore, less likely to be related to other alerts. Third, it can be used to anticipate the future steps of an attack and, thereby, come up with a strategy to reduce the damage.

References

Barbara D., Wu N., and Jajodia S., Detecting novel network intrusions using bayes estimators. In Proc. First SIAM Conference on Data Mining, Chicago, IL, April 2001.

Barbara D., Couto J., Jajodia S., and Wu N., Adam: Detecting Intrusions by Data Mining, In Proc. 2^{nd} Annual IEEE Information Assurance Workshop, West Point, NY, June 2001.

Cuppens F. and Miege A., Alert Correlation in a Cooperative Intrusion Detection Framework, Proc. IEEE Symposium on Security and Privacy, May 2002.

Ertoz L., Eilertson E., Lazarevic A., Tan P., Dokes P., Kumar V., Srivastava J., Detection of Novel Attacks using Data Mining, Proc. IEEE Workshop on Data Mining and Computer Security, November 2003.

Han J. and Kamber M., Data Mining: Concepts and Techniques, Morgan Kaufmann, August 2000.

Kumar V., Lazarevic A., Ertoz L., Ozgur A., Srivastava J., A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection, In Proc. Third SIAM International Conference on Data Mining, San Francisco, May 2003.

Lee W., Stolfo, S. J., and Kwok K. W. Mining audit data to build intrusion detection models. In Proc. Fourth International Conference on Knowledge Discovery and Data Mining, NewYork, 1998.

Lee W. and Stolfo S. J. Data Mining approaches for intrusion detection, In Proc. Seventh USENIX Security Symposium, San Antonio, TX, 1998.

Ning P., Cui Y., Reeves D. S., Constructing Attack Scenarios through Correlation of Intrusion Alerts, Proc. ACM Computer and Communications Security Conf., 2002.

Ning P., Xu D., earning Attack Strategies from Intrusion Alerts, Proc. ACM Computer and Communications Security Conf., 2003.

Portnoy L., Eskin E., Stolfo S. J., Intrusion Detection with unlabeled data using clustering. In Proceedings of ACM Workshop on Data Mining Applied to Security, 2001.