Chapter 1

DATA MINING FOR INTRUSION DETECTION

A Critical Review

Klaus Julisch
IBM Research
Zurich Research Laboratory
kju@zurich.ibm.com

Abstract
Data mining techniques have been successfully applied in many different fields including marketing, manufacturing, process control, fraud detection, and network management. Over the past five years, a growing number of research projects have applied data mining to various problems in intrusion detection. This chapter surveys a representative cross section of these research efforts. Moreover, four characteristics of contemporary research are identified and discussed in a critical manner. Conclusions are drawn and directions for future research are suggested.

Note: This article is an excerpt of the original work published in D. Barbará and S. Jajodia, editors, Applications of Data Mining in Computer Security, Kluwer Academic Publisher, Boston, 2002.

1. Introduction

Intrusion detection is the process of monitoring and analyzing the events occurring in a computer system in order to detect signs of security problems (Bace, 2000). Over the past ten years, intrusion detection and other security technologies such as cryptography, authentication, and firewalls have increasingly gained in importance (Allen et al., 2000). However, intrusion detection is not yet a perfect technology (Lippmann et al., 2000; Allen et al., 2000). This has given data mining the opportunity to make several important contributions to the field of intrusion detection (cf. Section 3).

This chapter gives a critical account of the past five years of data mining research in intrusion detection. To this end, we begin by introducing
data mining basics in Section 2. Section 3 surveys a representative selection of research projects that used data mining to address problems in intrusion detection. In Section 4, we identify and discuss four characteristics of contemporary and past research efforts. This discussion leads to Section 5, where we suggest new directions for future research. Section 6 summarizes the chapter.

We have attempted to make this chapter as self-contained as possible. However, given the interdisciplinary nature of the topic, it was not possible to write complete introductions to both, intrusion detection and data mining. We assumed that the reader has an intrusion detection background, and consequently put more emphasis on data mining basics. Complementary to this chapter, there is an abundance of excellent introductory material to both intrusion detection (Bace, 2000; Allen et al., 2000; Debar et al., 2000) as well as data mining (Han and Kamber, 2000; Mannila et al., 2001; Berry and Linoff, 1997) that can be consulted if needed.

2. Data Mining Basics

Historically, the notion of finding useful patterns in data has been given a variety of names including data mining, knowledge discovery in databases, information harvesting, data archaeology, and data pattern analysis (Fayyad et al., 1996a; Han and Kamber, 2000). Moreover, there has been some confusion about how data mining relates to the fields machine learning and statistics (Mannila, 1996). In Subsection 2.1, we clarify the terminology and the link to related fields. Section 2.2 describes four well-known data mining techniques that have been extensively used in intrusion detection. Section 2.3 concludes the discussion by summarizing several open research challenges in the field of data mining.

2.1 Data Mining, KDD, and Related Fields

The term data mining is frequently used to designate the process of extracting useful information from large databases. In this chapter, we adopt a slightly different view, which is identical to the one expressed by Fayyad et al. (1996b, Chapter 1) \(^1\). In this view, the term knowledge discovery in databases (KDD) is used to denote the process of extracting useful knowledge from large data sets. Data mining, by contrast, refers to one particular step in this process. Specifically, the data mining step applies so-called data mining techniques to extract patterns from the data. Additionally, it is preceded and followed by other KDD steps, which ensure that the extracted patterns actually correspond to
useful knowledge. Indeed, without these additional KDD steps, there is a high risk of finding meaningless or uninteresting patterns (Fayyad, 1998; Klemettinen et al., 1997; Stedman, 1997).

In other words, the KDD process uses data mining techniques along with any required pre- and post-processing to extract high-level knowledge from low-level data. In practice, the KDD process is interactive and iterative, involving numerous steps with many decisions being made by the user (Fayyad et al., 1996b, Chapter 2). Here, we broadly outline some of the most basic KDD steps:

1. **Understanding the application domain:** First is developing an understanding of the application domain, the relevant background knowledge, and the specific goals of the KDD endeavor.

2. **Data integration and selection:** Second is the integration of multiple (potentially heterogeneous) data sources and the selection of the subset of data that is relevant to the analysis task.

3. **Data mining:** Third is the application of specific algorithms for extracting patterns from data.

4. **Pattern evaluation:** Fourth is the interpretation and validation of the discovered patterns. The goal of this step is to guarantee that actual knowledge is being discovered.

5. **Knowledge representation:** This step involves documenting and using the discovered knowledge.

We next turn to the link between data mining and the related disciplines of machine learning and statistics. To begin with, data mining extensively uses known techniques from machine learning, statistics, and other fields. Nevertheless, several differences between data mining and related fields have been identified in the literature (Mannila, 1996; Glymour et al., 1997; Fayyad et al., 1996a). Specifically, one of the most frequently cited characteristics of data mining is its focus on finding relatively simple, but interpretable models in an efficient and scalable manner.

In other words, data mining emphasizes the efficient discovery of simple, but understandable models that can be interpreted as interesting or useful knowledge. Thus, for example, neural networks — although a powerful modeling tool — are relatively difficult to understand compared to rules (Cohen, 1995), trees (Quinlan, 1986), sequential patterns (Rigoutsos and Floratos, 1998), or associations (Agrawal et al., 1993). As a consequence, neural networks are of less practical importance in
data mining. This should not come as a surprise. In fact, data mining is just a step in the KDD process. As such, it has to contribute to the overall goal of knowledge discovery. Clearly, only understandable patterns can qualify as “knowledge”. Hence the importance of understandability in data mining.

### 2.2 Some Data Mining Techniques

Data mining techniques essentially are pattern discovery algorithms. Some techniques such as association rules (Agrawal et al., 1993) are unique to data mining, but most are drawn from related fields such as machine learning or pattern recognition. In this section, we introduce four well-known data mining techniques that have been widely used in intrusion detection. A broader and more detailed treatment of data mining techniques can be found elsewhere (Han and Kamber, 2000; Mannila et al., 2001; Berry and Linoff, 1997).

A potential source of confusion is that different data mining techniques assume different input data representations. For example, association rules have historically been discussed under the assumption that the input data is represented as a set of transactions (Agrawal et al., 1993; Agrawal and Srikant, 1994). Later, association rule mining over relational databases has been investigated (Srikant and Agrawal, 1996; Miller and Yang, 1997). Depending on the input data representations (sets of transactions versus relational databases), the association rule concept is presented differently. A related problem is that there are many different ways to represent the same data set in a relational database (Elmasri and Navathe, 1994). So one might wonder whether all these representations are equally adequate for the purpose of data mining. To avoid issues of data representation, we next define a unified input data format, for which all subsequent data mining techniques will be described. In practice, the available input data does not necessarily follow this format. Then, it is the responsibility of the second KDD step (“Data integration and selection”, as defined on page 3) to transform the available data into the format required by the data mining techniques.

#### 2.2.1 Association Rules

[...]

#### 2.2.2 Frequent Episode Rules

[...]

#### 2.2.3 Classification

[...]

#### 2.2.4 Clustering

[...]
2.3 Research Challenges in Data Mining

In a recent paper, Smyth (2001) has identified research challenges in data mining. Three years earlier, a similar list had been compiled by different authors (Grossman et al., 1998). In this section, we summarize the subset of the research challenges that are of direct relevance to intrusion detection:

3. Data Mining Meets Intrusion Detection

The goal of intrusion detection is to detect security violations in information systems. Intrusion detection is a passive approach to security as it monitors information systems and raises alarms when security violations are detected. Examples of security violations include the abuse of privileges or the use of attacks to exploit software or protocol vulnerabilities.

Traditionally, intrusion detection techniques are classified into two broad categories: misuse detection and anomaly detection (Mounjí, 1997, Chapter 2). Misuse detection works by searching for the traces or patterns of well-known attacks. Clearly, only known attacks that leave characteristic traces can be detected that way. Anomaly detection, on the other hand, uses a model of normal user or system behavior and flags significant deviations from this model as potentially malicious. This model of normal user or system behavior is commonly known as the user or system profile. A strength of anomaly detection is its ability to detect previously unknown attacks.

Additionally, intrusion detection systems (IDSs) are categorized according to the kind of input information they analyze. This leads to the distinction between host-based and network-based IDSs. Host-based IDSs analyze host-bound audit sources such as operating system audit trails, system logs, or application logs. Network-based IDSs analyze network packets that are captured on a network. More information on intrusion detection in general can be found, for example, in a recent book by Bace (2000).

In the past five years, a growing number of research projects have applied data mining to intrusion detection. Here, we survey a representative cross section of these projects. The intention of this survey is to give the reader a broad overview of the work that has been done at the intersection between intrusion detection and data mining. As a consequence, this section includes the most prominent projects in the field as well as some interesting niche projects that pursue less known
avenues. Specifically, our rationale for including the various projects into this survey is as follows:

- MADAM ID (cf. Section 3.1) is one of the first and, with almost a dozen conference and journal papers, certainly one of the best-known data mining projects in intrusion detection.
- In our eyes, ADAM (cf. Section 3.2) is the second most widely known and well-published project in the field.
- The clustering project of Section 3.3 is still very young and probably less known, but represents a novel and interesting research thrust.
- All of the above projects perform data mining on raw network data. In Section 3.4, we present three projects that apply data mining to intrusion detection alarms. This will broaden and balance our overview of the field.
- Section 3.5 rounds off this review and briefly mentions some of the other projects that we could not discuss in more detail.

3.1 MADAM ID

[...]  

3.2 ADAM  

[...]  

3.3 Clustering of Unlabeled ID Data  

[...]  

3.4 Mining the Alarm Stream  

[...]  

3.5 Further Reading

In this section, we briefly survey other relevant work that has not yet been mentioned². Wisdom & Sense (Vaccaro and Liepins, 1989) is probably the earliest system that can be considered as being based on data mining. Wisdom & Sense is an anomaly detection system that mines association rules from historical audit data to represent normal behavior. Similarly, Teng et al. (1990) use a form of automatically learned frequent episode rules to represent normal user behavior. The idea of Lankewicz
and Benard (1991) is to cluster audit log records and to represent each cluster by a single “typical” audit record. These typical audit records form the model of normal behavior against which future audit records are compared. A similar idea has been pursued by Lane and Brodley (1999), who cluster attack-free shell command sequences and define the “cluster centers” to represent normal behavior. Subsequently, anomalous command sequences can be detected based on their distance to the cluster centers. Mukkamala et al. (1999) use data mining techniques to reduce the amount of audit data that needs to be maintained and analyzed for intrusion detection. Similar work in audit data reduction has been reported by Lam et al. (1996). Finally, there is a long list of research projects that have tried to model system call sequences by a variety of different models, including neural networks, hidden Markov models, as well as fixed and variable length patterns. The work by Warrender et al. (1999) and Debar et al. (1998) is representative of this thrust of research.

4. Observations on the State of the Art

This section makes the following four observations about contemporary data mining efforts in intrusion detection:

- Most research concentrates on the construction of operational IDSs, rather than on the discovery of new and fundamental insights into the nature of attacks and false positives.

- It is very common to focus on the data mining step, while the other KDD steps are largely ignored.

- Much research is based on strong assumptions that complicate practical application.

- Up to now, data mining in intrusion detection focuses on a small subset of the spectrum of possible applications.

In the following sections, these observations will be discussed in a critical manner.

4.1 Data Mining, but no Knowledge Discovery

[...]

4.2 Disregard of Other KDD Steps

[...]
4.3 Too Strong Assumptions

[...]

4.4 Narrow Scope of Research Activities

[...]

5. Future Research Directions

[...]

6. Summary

This chapter has reviewed the past five years of data mining in intrusion detection. Based on this review, we have made four observations about contemporary and past research efforts. Very briefly, we observed a focus on building operational IDSs, a disregard for the overall KDD process, the reliance on labeled high-quality training data, and the focus on a few, admittedly important problems. We have discussed these observations in a critical manner, which has lead us to the following recommendations for future research:

- Future projects should pay closer attention to the KDD process.
- Either more work should address the (semi-)automatic generation of high-quality labeled training data, or the existence of such data should no longer be assumed.
- Future projects should explore novel applications of data mining that do not fall into the categories feature selection and anomaly detection.
- To deal with some of the general challenges in data mining, it might be best to develop special-purpose solutions that are tailored to intrusion detection.

Acknowledgments

The author thanks Birgit Baum-Waidner, Marc Dacier, Andreas Wespi, and Diego Zamboni for their valuable comments on earlier versions of this chapter.

This research was supported by the European IST Project MAFTIA (IST-1999-11583), which is partially funded by the European Commission and the Swiss Federal Office for Education and Science. The views herein are those of the author and do not necessarily reflect the views of the supporting agencies.
Notes

1. Other authors also support this view (Mannila, 1996; Han and Kamber, 2000).
2. To avoid redundancy, we deliberately refrain from discussing the other papers in this volume.

References


REFERENCES


Interestingness in Knowledge Discovery. In Proceedings of the First
International Conference on Knowledge Discovery and Data Mining,
pages 275–281.
cceedings of the 4th IEEE Aerospace Computer Security Applications
Conference, Orlando, FL, pages 37–44.
in Data Mining. In Proceedings of the ACM SIGMOD International
Workshop on Research Issues in Data Mining and Knowledge Discov-
ery (DMKD’01).
Rules in Large Relational Tables. In Proceedings of the 1996 ACM-
SIGMOD Conference on Management of Data, pages 1–12.
Detection Using Inductively Generated Sequential Patterns. In Pro-
ceedings of the IEEE Symposium on Research in Security and Privacy,
Oakland, CA, pages 278–284.
puter Session Activity. In Proceedings of the IEEE Symposium on
Warrender, C., Forrest, S., and Pearlmutter, B. (1999). Detecting Intru-
sions Using System Calls: Alternative Data Models. In Proceedings of
the IEEE Symposium on Research in Security and Privacy, Oakland,
CA, pages 133–145.