Data Mining and Machine Learning in Cybersecurity

Sumeet Dua and Xian Du

CRC Press
Taylor & Francis Group
AN AUERBACH BOOK

Data Mining and Machine Learning in Cybersecurity

Data Mining and Machine Learning in Cybersecurity

Sumeet Dua and Xian Du

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
AN AUERBACH BOOK

Auerbach Publications
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-1-4398-3943-0 (Ebook-PDF)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the Auerbach Web site at
http://www.auerbach-publications.com

Contents

List of Figures

List of Tables

Preface

In the emerging era of Web 3.0, securing cyberspace has gradually evolved into a critical organizational and national research agenda inviting interest from a multidisciplinary scientific workforce. There are many avenues into this area, and, in recent research, machine-learning and data-mining techniques have been applied to design, develop, and improve algorithms and frameworks for cybersecurity system design. Intellectual products in this domain have appeared under various topics, including machine learning, data mining, cybersecurity, data management and modeling, and privacy preservation. Several conferences, workshops, and journals focus on the fragmented research topics in this area. However, transcendent and interdisciplinary assessment of past and current works in the field and possible paths for future research in the area are essential for consistent research and development.

This interdisciplinary assessment is especially useful for students, who typically learn cybersecurity, machine learning, and data mining in independent courses. Machine learning and data mining play significant roles in cybersecurity, especially as more challenges appear with the rapid development of information discovery techniques, such as those originating from the sheer dimensionality and heterogeneous nature of the network data, the dynamic change of threats, and the severe imbalanced classes of normal and anomalous behaviors. In this book, we attempt to combine all the above knowledge for a single advanced course.

This book surveys cybersecurity problems and state-of-the-art machine-learning and data-mining solutions that address the overarching research problems, and it is designed for students and researchers studying or working on machine learning and data mining in cybersecurity applications. The inclusion of cybersecurity in machine-learning research is important for academic research. Such an inclusion inspires fundamental research in machine learning and data mining, such as research in the subfields of imbalanced learning, feature extraction for data with evolving characteristics, and privacy-preserving data mining.

Organization

In Chapter 1, we introduce the vulnerabilities of cyberinfrastructure and the conventional approaches to cyber defense. Then, we present the vulnerabilities of these conventional cyber protection methods and introduce higher-level methodologies that use advanced machine learning and data mining to build more reliable cyber defense systems. We review the cybersecurity solutions that use machine-learning and data-mining techniques, including privacy-preservation data mining, misuse detection, anomaly detection, hybrid detection, scan detection, and profiling detection. In addition, we list a number of references that address cybersecurity issues using machine-learning and data-mining technology to help readers access the related material easily.

In Chapter 2, we introduce machine-learning paradigms and cybersecurity along with a brief overview of machine-learning formulations and the application of machine-learning methods and data mining/management in cybersecurity. We discuss challenging problems and future research directions that are possible when machine-learning methods are applied to the huge amount of temporal and unbalanced network data.

In Chapter 3, we address misuse/signature detection. We introduce fundamental knowledge, key issues, and challenges in misuse/signature detection systems, such as building efficient rule-based algorithms, feature selection for rule matching and accuracy improvement, and supervised machine-learning classification of attack patterns. We investigate several supervised learning methods in misuse detection. We explore the limitations and difficulties of using these machine-learning methods in misuse detection systems and outline possible problems, such as the inadequate ability to detect a novel attack, irregular performance for different attack types, and requirements of the intelligent feature selection. We guide readers to questions and resources that will help them learn more about the use of advanced machine-learning techniques to solve these problems.

In Chapter 4, we provide an overview of anomaly detection techniques. We investigate and classify a large number of machine-learning methods in anomaly detection. In this chapter, we briefly describe the applications of machine-learning methods in anomaly detection. We focus on the limitations and difficulties that encumber machine-learning methods in anomaly detection systems. Such problems include an inadequate ability to maintain a high detection rate and a low false-alarm rate. As anomaly detection is the most concentrative application area of machine-learning methods, we perform in-depth studies to explain the appropriate learning procedures, e.g., feature selection, in detail.

In Chapter 5, we address hybrid intrusion detection techniques. We describe how hybrid detection methods are designed and employed to detect unknown intrusions and anomaly detection with a lower false-positive rate. We categorize the hybrid intrusion detection techniques into three groups based on combinational methods. We demonstrate several machine-learning hybrids that raise detection accuracies in the intrusion detection system, including correlation techniques, artificial neural networks, association rules, and random forest classifiers.

In Chapter 6, we address scan detection techniques using machine-learning methods. We explain the dynamics of scan attacks and focus on solving scan detection problems in applications. We provide several examples of machine-learning methods used for scan detection, including the rule-based methods, threshold random walk, association memory learning techniques, and expert knowledge-rule-based learning model. This chapter addresses the issues pertaining to the high percentage of false alarms and the evaluation of efficiency and effectiveness of scan detection.

In Chapter 7, we address machine-learning techniques for profiling network traffic. We illustrate a number of profiling modules that profile normal or anomalous behaviors in cyberinfrastructure for intrusion detection. We introduce and investigate a number of new concepts for clustering methods in intrusion detection systems, including association rules, shared nearest neighbor clustering, EM-based clustering, subspace, and informatics theoretic techniques. In this chapter, we address the difficulties of mining the huge amount of streaming data and the necessity of interpreting the profiling results in an understandable way.

In Chapter 8, we provide a comprehensive overview of available machine-learning technologies in privacy-preserving data mining. In this chapter, we concentrate on how data-mining techniques lead to privacy breach and how privacy-preserving data mining achieves data protection via machine-learning methods. Privacy-preserving data mining is a new area, and we hope to inspire research beyond the foundations of data mining and privacy-preserving data mining.

In Chapter 9, we describe the emerging challenges in fixed computing or mobile applications and existing and potential countermeasures using machine-learning methods in cybersecurity. We also explore how the emerging cyber threats may evolve in the future and what corresponding strategies can combat threats. We describe the emerging issues in network monitoring, profiling, and privacy preservation and the emerging challenges in intrusion detection, especially those challenges for anomaly detection systems.

Authors

Dr. Sumeet Dua is currently an Upchurch endowed associate professor and the coordinator of IT research at Louisiana Tech University, Ruston, Louisiana. He received his PhD in computer science from Louisiana State University, Baton Rouge, Louisiana.

His areas of expertise include data mining, image processing and computational decision support, pattern recognition, data warehousing, biomedical informatics, and heterogeneous distributed data integration. The National Science Foundation (NSF), the National Institutes of Health (NIH), the Air Force Research Laboratory (AFRL), the Air Force Office of Sponsored Research (AFOSR), the National Aeronautics and Space Administration (NASA), and the Louisiana Board of Regents (LA-BoR) have funded his research with over $2.8 million. He frequently serves as a study section member (expert panelist) for the National Institutes of Health (NIH) and panelist for the National Science Foundation (NSF)/CISE Directorate. Dr. Dua has chaired several conference sessions in the area of data mining and is the program chair for the *Fifth International Conference on Information Systems, Technology, and Management* (ICISTM-2011). He has given more than 26 invited talks on data mining and its applications at international academic and industry arenas, has advised more than 25 graduate theses, and currently advises several graduate students in the discipline. Dr. Dua is a coinventor of two issued U.S. patents, has (co-)authored more than 50 publications and book chapters, and has authored or edited four books. Dr. Dua has received the Engineering and Science Foundation Award for Faculty Excellence (2006) and the Faculty Research Recognition Award (2007), has been recognized as a distinguished researcher (2004–2010) by the Louisiana Biomedical Research Network (NIH-sponsored), and has won the Outstanding Poster Award at the NIH/NCI caBIG—NCRI Informatics Joint Conference; Biomedical Informatics without Borders: From Collaboration to Implementation. Dr. Dua is a senior member of the IEEE Computer Society, a senior member of the ACM, and a member of SPIE and the American Association for Advancement of Science.

Dr. Xian Du is a research associate and postdoctoral fellow at the Louisiana Tech University, Ruston, Louisiana. He worked as a postdoctoral researcher at the Centre National de la Recherche Scientifique (CNRS) in the CREATIS Lab, Lyon, France, from 2007 to 2008 and served as a software engineer in Kikuze Solutions Pte. Ltd., Singapore, in 2006. He received his PhD from the Singapore–MIT Alliance (SMA) Programme at the National University of Singapore in 2006.

Dr. Xian Du's current research focus is on high-performance computing using machine-learning and data-mining technologies, data-mining applications for cybersecurity, software in multiple computer operational environments, and clustering theoretical research. He has broad experience in machine-learning applications in industry and academic research at high-level research institutes. During his work in the CREATIS Lab in France, he developed a 3D smooth active contour technology for knee cartilage MRI image segmentation. He led a small research and development group to develop color control plug-ins for an RGB color printer to connect to the Windows® system through image processing GDI functions for Kikuze Solutions. He helped to build an intelligent e-diagnostics system for reducing mean time to repair wire-bonding machines at National Semiconductor Ltd., Singapore (NSC). During his PhD dissertation research at the SMA, he developed an intelligent color print process control system for color printers. Dr. Du's major research interests are machine-learning and data-mining applications, heterogeneous data integration and visualization, cybersecurity, and clustering theoretical research.

Chapter 1

Introduction

> Many of the nation's essential and emergency services, as well as our critical infrastructure, rely on the uninterrupted use of the Internet and the communications systems, data, monitoring, and control systems that comprise our cyber infrastructure. A cyber attack could be debilitating to our highly interdependent Critical Infrastructure and Key Resources (CIKR) and ultimately to our economy and national security.
>
> **Homeland Security Council**
> *National Strategy for Homeland Security, 2007*

The ubiquity of cyberinfrastructure facilitates beneficial activities through rapid information sharing and utilization, while its vulnerabilities generate opportunities for our adversaries to perform malicious activities within the infrastructure.* Because of these opportunities for malicious activities, nearly every aspect of cyberinfrastructure needs protection (Homeland Security Council, 2007).

Vulnerabilities in cyberinfrastructure can be attacked horizontally or vertically. Hence, cyber threats can be evaluated horizontally from the perspective of the attacker(s) or vertically from the perspective of the victims. First, we look at cyber threats vertically, from the perspective of the victims. A variety of adversarial agents such as nation-states, criminal organizations, terrorists, hackers, and other malicious users can compromise governmental homeland security through networks.

---

* Cyberinfrastructure consists of digital data, data flows, and the supportive hardware and software. The infrastructure is responsible for data collection, data transformation, traffic flow, data processing, privacy protection, and the supervision, administration, and control of working environments. For example, in our daily activities in cyberspace, we use health Supervisory Control and Data Acquisition (SCADA) systems and the Internet (Chandola et al., 2009).

For example, hackers may utilize personal computers remotely to conspire, proselytize, recruit accomplices, raise funds, and collude during ongoing attacks. Adversarial governments and agencies can launch cyber attacks on the hardware and software of the opponents' cyberinfrastructures by supporting financially and technically malicious network exploitations.

Cyber criminals threaten financial infrastructures, and they could pose threats to national economies if recruited by the adversarial agents or terrorist organizations. Similarly, private organizations, e.g., banks, must protect confidential business or private information from such hackers. For example, the disclosure of business or private financial data to cyber criminals can lead to financial loss via Internet banking and related online resources. In the pharmaceutical industry, disclosure of protected company information can benefit competitors and lead to market-share loss. Individuals must also be vigilant against cyber crimes and malicious use of Internet technology.

As technology has improved, users have become more tech savvy. People communicate and cooperate efficiently through networks, such as the Internet, which are facilitated by the rapid development of digital information technologies, such as personal computers and personal digital assistants (PDAs). Through these digital devices linked by the Internet, hackers also attack personal privacy using a variety of weapons, such as viruses, Trojans, worms, botnet attacks, rootkits, adware, spam, and social engineering platforms.

Next, we look at cyber threats horizontally from the perspective of the victims. We consider any malicious activity in cyberspace as a cyber threat. A cyber threat may result in the loss of or damage to cyber components or physical resources. Most cyber threats are categorized into one of three groups according to the intruder's purpose: stealing confidential information, manipulating the components of cyberinfrastructure, and/or denying the functions of the infrastructure. If we evaluate cyber threats horizontally, we can investigate cyber threats and the subsequent problems. We will focus on intentional cyber crimes and will not address breaches caused by normal users through unintentional operations, such as errors and omissions, since education and proper habits could help to avoid these threats.* We also will not explain cyber threats caused by natural disasters, such as accidental breaches caused by earthquakes, storms, or hurricanes, as these threats happen suddenly and are beyond our control.

1.1 Cybersecurity

To secure cyberinfrastructure against intentional and potentially malicious threats, a growing collaborative effort between cybersecurity professionals and researchers from institutions, private industries, academia, and government agencies has engaged in

* We define a normal cyber user as an individual or group of individuals who do not intend to intrude on the cybersecurity of other individuals.

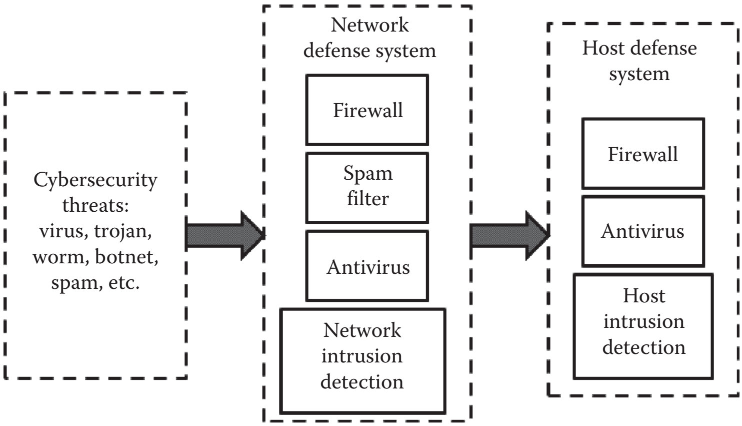

Figure 1.1 Conventional cybersecurity system.

exploiting and designing a variety of cyber defense systems. Cybersecurity researchers and designers aim to maintain the confidentiality, integrity, and availability of information and information management systems through various cyber defense systems that protect computers and networks from hackers who may want to intrude on a system or steal financial, medical, or other identity-based information.*

As shown in Figure 1.1, conventional cybersecurity systems address various cybersecurity threats, including viruses, Trojans, worms, spam, and botnets. These cybersecurity systems combat cybersecurity threats at two levels and provide network- and host-based defenses. Network-based defense systems control network flow by network firewall, spam filter, antivirus, and network intrusion detection techniques. Host-based defense systems control upcoming data in a workstation by firewall, antivirus, and intrusion detection techniques installed in hosts.

Conventional approaches to cyber defense are mechanisms designed in firewalls, authentication tools, and network servers that monitor, track, and block viruses and other malicious cyber attacks. For example, the Microsoft Windows® operating system has a built-in Kerberos cryptography system that protects user information. Antivirus software is designed and installed in personal computers and cyberinfrastructures to ensure customer information is not used maliciously. These approaches create a protective shield for cyberinfrastructure.

However, the vulnerabilities of these methods are ubiquitous in applications because of the flawed design and implementation of software and network

* The three requirements of cybersecurity correspond to the three types of intentional threats: confidentiality signifies the ability to prevent sensitive data from being disclosed to third parties; integrity ensures the infrastructure is complete and accurate, and availability refers to the accessibility of the normal operations of cyberinfrastructures, such as delivering and storing data.

infrastructure. Patches have been developed to protect the cyber systems, but attackers continuously exploit newly discovered flaws. Because of the constantly evolving cyber threats, building defense systems for discovered attacks is not enough to protect users. Higher-level methodologies are also required to discover the embedded and lurking cyber intrusions and cyber intrusion techniques, so that a more reliable security cyberinfrastructure can be utilized.

Many higher-level adaptive cyber defense systems can be partitioned into components as shown in Figure 1.2. Figure 1.2 outlines the five-step process for those defense systems. We discuss each step below.

Data-capturing tools, such as Libpcap for Linux®, Solaris BSM for SUN®, and Winpcap for Windows®, capture events from the audit trails of resource information sources (e.g., network). Events can be host-based or network-based depending on where they originate. If an event originates with log files, then it is categorized as a host-based event. If it originates with network traffic, then it is categorized as a network-based event. A host-based event includes a sequence of commands executed by a user and a sequence of system calls launched by an application, e.g., send mail. A network-based event includes network traffic data, e.g., a sequence of internet protocol (IP) or transmission control protocol (TCP) network packets. The data-preprocessing module filters out the attacks for which good signatures have been learned.

A feature extractor derives basic features that are useful in event analysis engines, including a sequence of system calls, start time, duration of a network flow, source IP and source port, destination IP and destination port, protocol,

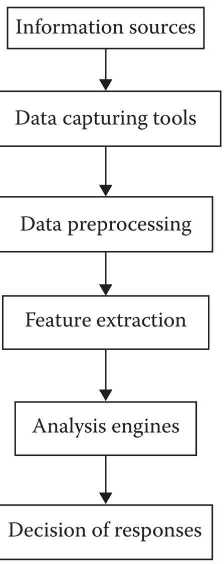

Figure 1.2 Adaptive defense system for cybersecurity.

number of bytes, and number of packets. In an analysis engine, various intrusion detection methods are implemented to investigate the behavior of the cyberinfrastructure, which may or may not have appeared before in the record, e.g., to detect anomalous traffic. The decision of responses is deployed once a cyber attack is identified. As shown in Figure 1.2, analysis engines are the core technologies for the generation of the adaptation ability of the cyber defense system. As discussed above, the solutions to cybersecurity problems include proactive and reactive security solutions.

Proactive approaches anticipate and eliminate vulnerabilities in the cyber system, while remaining prepared to defend effectively and rapidly against attacks. To function correctly, proactive security solutions require user authentication (e.g., user password and biometrics), a system capable of avoiding programming errors, and information protection [e.g., privacy-preserving data mining (PPDM)]. PPDM protects data from being explored by data-mining techniques in cybersecurity applications. We will discuss this technique in detail in Chapter 8. Proactive approaches have been used as the first line of defense against cybersecurity breaches. It is not possible to build a system that has no security vulnerabilities. Vulnerabilities in common security components, such as firewalls, are inevitable due to design and programming errors.

The second line of cyber defense is composed of reactive security solutions, such as intrusion detection systems (IDSs). IDSs detect intrusions based on the information from log files and network flow, so that the extent of damage can be determined, hackers can be tracked down, and similar attacks can be prevented in the future.

1.2 Data Mining

Due to the availability of large amounts of data in cyberinfrastructure and the number of cyber criminals attempting to gain access to the data, data mining, machine learning, statistics, and other interdisciplinary capabilities are needed to address the challenges of cybersecurity. Because IDSs use data mining and machine learning, we will focus on these areas. Data mining is the extraction, or "mining," of knowledge from a large amount of data. The strong patterns or rules detected by data-mining techniques can be used for the nontrivial prediction of new data. In nontrivial prediction, information that is implicitly presented in the data, but was previously unknown is discovered. Data-mining techniques use statistics, artificial intelligence, and pattern recognition of data in order to group or extract behaviors or entities. Thus, data mining is an interdisciplinary field that employs the use of analysis tools from statistical models, mathematical algorithms, and machine-learning methods to discover previously unknown, valid patterns and relationships in large data sets, which are useful for finding hackers and preserving privacy in cybersecurity.

Data mining is used in many domains, including finance, engineering, biomedicine, and cybersecurity. There are two categories of data-mining methods: supervised and unsupervised. Supervised data-mining techniques predict a hidden function using training data. The training data have pairs of input variables and output labels or classes. The output of the method can predict a class label of the input variables. Examples of supervised mining are classification and prediction. Unsupervised data mining is an attempt to identify hidden patterns from given data without introducing training data (i.e., pairs of input and class labels). Typical examples of unsupervised mining are clustering and associative rule mining.

Data mining is also an integral part of knowledge discovery in databases (KDDs), an iterative process of the nontrivial extraction of information from data and can be applied to developing secure cyberinfrastructures. KDD includes several steps from the collection of raw data to the creation of new knowledge. The iterative process consists of the following steps: data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and knowledge representation, as described below.

Step 1. During data cleaning, which is also known as data cleansing, noise and irrelevant data are removed from the collection.

Step 2. Data integration combines data from multiple and heterogeneous sources into one database.

Step 3. Data-selection techniques allow the user to obtain a reduced representation of the data set to keep the integrity of the original data set in a reduced volume.

Step 4. In data transformation, the selected data is transformed into suitable formats.

Step 5. Data mining is the stage in which analysis tools are applied to discover potentially useful patterns.

Step 6. Pattern evaluation identifies interesting and useful patterns using given validation measures.

Step 7. In knowledge representation, the final phase of the knowledge-discovery process, discovered knowledge is presented to the users in visual forms.

Data-mining techniques are used to aid in the development of predictive models that enable a real-time cyber response after a sequence of cybersecurity processes, which include real-time data sampling, selection, analysis and query, and mining peta-scale data to classify and detect attacks and intrusions on a computer network (Denning, 1987; Lee and Stolfo, 1998; Axelsson, 2000; Chandola et al., 2006; Homeland Security Council, 2007). Learning user patterns and/or behaviors is critical for intrusion detection and attack predictions. Learning these behaviors is important, as they can identify and describe structural patterns in the data automatically and theoretically explain data and predict patterns. Automatic and theoretic learning require complex computation that calls for abundant machine-learning algorithms. We will discuss the concept of machine learning in Section 1.3.

1.3 Machine Learning

Learning is the process of building a scientific model after discovering knowledge from a sample data set or data sets. Generally, machine learning is considered to be the process of applying a computing-based resource to implement learning algorithms. Formally, machine learning is defined as the complex computation process of automatic pattern recognition and intelligent decision making based on training sample data.

Machine-learning methods can be categorized into four groups of learning activities: symbol-based, connectionist-based, behavior-based, and immune system-based activities. Symbol-based machine learning has a hypothesis that all knowledge can be represented in symbols and that machine learning can create new symbols and new knowledge, based on the known symbols. In symbol-based machine learning, decisions are deducted using logical inference procedures. Connectionist-based machine learning is constructed by imitating neuron net connection systems in the brain. In connectionist machine learning, decisions are made after the systems are trained and patterns are recognized. Behavior-based learning has the assumption that there are solutions to behavior identification, and is designed to find the best solution to solve the problem. The immune-system-based approach learns from its encounters with foreign objects and develops the ability to indentify patterns in data. None of these machine-learning methods has noticeable advantages over the others. Thus, it is not necessary to select machine-learning methods based on these fundamental distinctions, and within the machine-learning process, mathematical models are built to describe the data randomly sampled from an unseen probability distribution.

Machine learning has to be evaluated empirically because its performance heavily depends on the type of training experience the learning machine has undergone, the performance evaluation metrics, and the strength of the problem definition. Machine-learning methods are evaluated by comparing the learning results of methods applied on the same data set or quantifying the learning results of the same methods applied on sample data sets. The measure metrics will be discussed in Section 2.2.4. In addition to the accuracy evaluation, the time complexity and feasibility of machine learning are studied (Debar et al., 1999). Generally, the feasibility of a machine-learning method is acceptable when its computation time is polynomial.

Machine-learning methods use training patterns to learn or estimate the form of a classifier model. The models can be parametric or unparametric. The goal of using machine-learning algorithms is to reduce the classification error on the given training sample data. The training data are finite such that the learning theory requires probability bounds on the performance of learning algorithms. Depending on the availability of training data and the desired outcome of the learning algorithms, machine-learning algorithms are categorized into supervised learning and unsupervised learning. The first two groups include most machine-learning applications in cybersecurity. In supervised learning, pairs of input and target output are given to train a function, and a learning model is trained such that the output of the function can be predicted at a minimum cost. The supervised learning methods are categorized based on the structures and objective functions of learning algorithms. Popular categorizations include artificial neural network (ANN), support vector machine (SVM), and decision trees.

In unsupervised learning, no target or label is given in sample data. Unsupervised learning methods are designed to summarize the key features of the data and to form the natural clusters of input patterns given a particular cost function. The most famous unsupervised learning methods include *k*-means clustering, hierarchical clustering, and self-organization map. Unsupervised learning is difficult to evaluate, because it does not have an explicit teacher and, thus, does not have labeled data for testing.

We will discuss a number of classic machine-learning methods in Chapter 2. Readers who are familiar with this topic may skip that material.

1.4 Review of Cybersecurity Solutions

A number of surveys and review articles have focused on intrusion detection technologies (Debar et al., 1999; Axelsson, 2000; Homeland Security Council, 2007; Patcha and Park, 2007) or data mining in specific applications (Stolfo et al., 2001; Chandola et al., 2006). Hodge and Austin (2004) categorized anomaly detection techniques in statistics, neural networks, machine learning, and hybrid approaches. Meza et al. (2009) highlighted important cybersecurity problems such as cybersecurity for mathematical and statistical solutions. Siddiqui et al. (2008) categorized data-mining techniques for malware detection based on file features and analysis (static or dynamic) and detection types. Lee and Fan (2001) described a data-mining framework for mining audit data using IDSs.

In Section 1.4.1, we provide a broad structural review of the uses of machine learning for data mining in cybersecurity in the past 10 years. Besides the traditional intrusion detection (adaptive defense system) technologies, we also review proactive cybersecurity solutions. We focus on PPDM, which is designed to protect data from being explored by machine learning for data mining in cybersecurity applications. Scan detection, profiling, and hybrid detection are added to the traditional misuse and anomaly detection technologies in reactive security solutions.

1.4.1 Proactive Security Solutions

Traditionally, proactive security solutions (Canetti et al., 1997; Barak et al., 1999) are designed to maintain the overall security of a system, even if individual components of the system have been compromised by an attack.

Recently, the improvement of data-mining techniques and information technology brings unlimited chances for Internet and other media users to explore new information. The new information may include sensitive information and, thus, incur a new research domain where researchers consider data-mining algorithms from the viewpoint of privacy preservation. This new research, called PPDM

Table 1.1 Examples of PPDM

Data-Mining Techniques	*Privacy-Preservation Methods*	*References*
A.1 Statistical methods	Heuristic-based	Du et al. (2004)
A.2 Bayesian networks (BNs)	Reconstruction-based	Wright and Yang (2004)
A.3 Unsupervised clustering algorithm	Heuristic-based	Vaidya and Clifton (2003)
A.4 Association rules	Reconstruction-based	Evfimievski et al. (2002)
A.5 ANNs	Cryptography-based	Barni et al. (2006)
A.6 Decision tree	Cryptography-based	Du and Zhan (2002), Agrawal and Srikant (2000)
A.7 *k*-nearest neighbor (KNN)	Cryptography-based	Kantarcioglu and Clifton (2004)
A.8 SVM	Reconstruction-based	Yu et al. (2006)

Note: The privacy-preservation techniques, the most important techniques for the selective modification of the data, are categorized into three groups: heuristic-based techniques, cryptography-based techniques, and reconstruction-based techniques (see details in Verykios et al., 2004).

(Agrawal and Srikant, 2000; Verykios et al., 2004), is designed to protect private data and knowledge in data mining. PPDM methods can be characterized by data distribution, data modification, data-mining algorithms, rule hiding, and privacy-preservation techniques. We categorize the principle PPDM methods in Table 1.1 according to machine-learning algorithms for data mining and present their privacy-preservation methods. We discuss these methods in Chapter 8.

At this point in its research history, PPDM algorithms are developed for individual various machine-learning methods. The PPDM algorithms include privacy-preserving decision tree (Chebrolu et al., 2005), privacy-preserving association rule mining (Evfimievski et al., 2002), privacy-preserving clustering (Vaidya and Clifton, 2003), and privacy-preserving SVM classification (Yu et al., 2006) (see Table 1.1). We address PPDM and its application studies in Chapter 8.

1.4.2 Reactive Security Solutions

Since the principles of intrusion detection were first introduced by Denning in 1987, large numbers of reactive security systems have been developed. Such systems include RIPPER (Lee and Stolfo, 2000), EMERALD (Porras and Neumann, 1997), MADAM ID (Lee and Stolfo, 2000), LERAD (Mahoney and Chan, 2002), and MINDS (Chandola et al., 2006).

Cyber intrusion is defined as any unauthorized attempt to access, manipulate, modify, or destroy information or to use a computer system remotely to spam, hack, or modify other computers. An IDS intelligently monitors activities that occur in a computing resource, e.g., network traffic and computer usage, to analyze the events and to generate reactions. In IDSs, it is always assumed that an intrusion will manifest itself in a trace of these events, and the trace of an intrusion is different from traces left by normal behaviors. To achieve this purpose, network packets are collected, and the rule violation is checked with pattern recognition methods. An IDS system usually monitors and analyzes user and system activities, accesses the integrity of the system and data, recognizes malicious activity patterns, generates reactions to intrusions, and reports the outcome of detection.

The activities that the IDSs trace can form a variety of patterns or come from a variety of sources. According to the detection principles, we classify intrusion detection into the following modules: misuse/signature detection, anomaly detection algorithms, hybrid detection, and scan detector and profiling modules. Furthermore, IDSs recognize and prevent malicious activities through network- or host-based methods. These IDSs search for specific malicious patterns to identify the underlying suspicious intent. When an IDS searches for malicious patterns in network traffic, we call it a network-based IDS. When an IDS searches for malicious patterns in log files, we call it host-based IDS.

1.4.2.1 Misuse/Signature Detection

Misuse detection, also called signature detection, is an IDS triggering method that generates alarms when a known cyber misuse occurs. A signature detection technique measures the similarity between input events and the signatures of known intrusions. It flags behavior that shares similarities with a predefined pattern of intrusion. Thus, known attacks can be detected immediately and realizably with a lower false-positive rate. However, signature detection cannot detect novel attacks. Examples of data mining in misuse detection are listed in Table 1.2. We address misuse detection techniques in Chapter 3.

1.4.2.2 Anomaly Detection

Anomaly detection triggers alarms when the detected object behaves significantly differently from the predefined normal patterns. Hence, anomaly detection techniques are designed to detect patterns that deviate from an expected normal model built for the data. In cybersecurity, anomaly detection includes the detection of malicious activities, e.g., penetrations and denial of service. The approach consists of two steps: training and detection. In the training step, machine-learning

Table 1.2 Examples of Data Mining and Machine Learning for Misuse/Signature Detection

Technique Used	*Input Data Format*	*Levels*	*References*
B.1 Rule-based signature analysis	Frequency of system calls, off line	Host	Lee et al. (1999)
B.2 ANN	TCP/IP data, offline	Host	Ghosh and Schwartzbard (1999), Cannady (1998)
B.3 Fuzzy association rules	Frequency of system calls, online	Host	Abraham et al. (2007b), Su et al. (2009)
B.4 SVM	TCP/IP data, offline	Network	Mukkamala and Sung (2003)
B.5 Linear genetic programs (LGP)	TCP/IP data, offline	Network	Mukkamala and Sung (2003), Abraham et al. (2007a,b), Srinivas et al. (2004)
B.6 Classification and regression trees	Frequency of system calls, offline	Host	Chebrolu et al. (2005)
B.7 Decision tree	TCP/IP data, online	Network	Kruegel and Toth (2003)
B.8 BN	Frequency of system calls, offline	Host	Chebrolu et al. (2005)
B.9 Statistical method	Executables, offline	Host	Schultz et al. (2001)

techniques are applied to generate a profile of normal patterns in the absence of an attack. In the detection step, the input events are labeled as attacks if the event records deviate significantly from the normal profile. Subsequently, anomaly detection can detect previously unknown attacks. However, anomaly detection is hampered by a high rate of false alarms. Moreover, the selection of inappropriate features can hurt the effectiveness of the detection result, which corresponds to the learned patterns. In extreme cases, a malicious user can use anomaly data as normal data to train an anomaly detection system, so that it will recognize malicious patterns as normal. Examples of data mining in anomaly detection are listed in Table 1.3. We will address anomaly detection techniques in Chapter 4.

Table 1.3 Examples of Data Mining and Machine Learning for Anomaly Detection

Technique Used	*Input Data Format*	*Levels*	*References*
C.1 Statistical methods	Sequences of system calls, offline	Host	Ye et al. (2001), Feinstein et al. (2003), Smaha (1988), Ye et al. (2002)
C.2 Statistical methods	TCP/IP data, online	Network	Yamanishi and Takeuchi (2001), Yamanishi et al. (2000), Mahoney and Chan (2002, 2003), Soule et al. (2005)
C.3 Unsupervised clustering algorithm	TCP/IP data, offline	Network	Portnoy et al. (2001), Leung and Leckie (2005), Warrender et al. (1999), Zhang and Zulkernine (2006a,b)
C.4 Subspace	TCP/IP data offline	Network	Li et al. (2006)
C.5 Information theoretic	TCP/IP, online	Network	Lakhina et al. (2005)
C.6 Association rules	Frequency of system calls, online	Host	Lee and Stolfo (1998), Abraham et al. (2007a,b), Su et al. (2009), Lee et al. (1999)
C.7 Kalman filter	TCP/IP data, online	Network	Soule et al. (2005)
C.8 Hidden Markov model (HMM)	Sequences of system calls, offline	Host	Warrender et al. (1999)
C.9 ANN	Sequences of system calls, offline	Host	Ghosh et al. (1998, 1999), Liu et al. (2002)
C.10 Principal component analysis (PCA)	TCP/IP data, online	Network	Lakhina et al. (2004), Ringberg et al. (2007)
C.11 KNN	Frequency of system calls, offline	Host	Liao and Vemuri (2002)
C.12 SVM	TCP/IP data, offline	Network	Hu et al. (2003), Chen et al. (2005)

Table 1.4 Examples of Data Mining for Hybrid Intrusion Detection

Technique Used	*Input Data Format*	*Levels*	*References*
D.1 Correlation	TCP/IP data, online	Network	Ning et al. (2004), Cuppens and Miège (2002), Dain and Cunningham (2001a,b)
D.2 Statistical methods	Sequences of system calls, offline	Host	Endler (1998)
D.3 ANN	Sequences of system calls, offline	Host	Endler (1998)
D.4 Association rules	Frequency of system calls, online	Host	Lee and Stolfo (2000)
D.5 ANN	TCP/IP data, online	Network	Ghosh et al. (1999)
D.6 Random forest	TCP/IP data, online	Network	Zhang and Zulkernine (2006a,b)

1.4.2.3 Hybrid Detection

Most current IDSs employ either misuse detection techniques or anomaly detection techniques. Both of these methods have drawbacks: misuse detection techniques lack the ability to detect unknown intrusions; anomaly detection techniques usually produce a high percentage of false alarms. To improve the techniques of IDSs, researchers have proposed hybrid detection techniques to combine anomaly and misuse detection techniques in IDSs. Examples for hybrid detection techniques are listed in Table 1.4. We address hybrid detection techniques in Chapter 5.

1.4.2.4 Scan Detection

Scan detection generates alerts when attackers scan services or computer components in network systems before launching attacks. A scan detector identifies the precursor of an attack on a network, e.g., destination IPs and the source IPs of Internet connections. Although many scan detection techniques have been proposed and declared to be able to detect the precursors of cyber attacks, the high false-positive rate or the low scan detection rate limits the application of these solutions in practice. Some examples of scan detection techniques are categorized in Table 1.5. We address scan and scan detection techniques in Chapter 6.

1.4.2.5 Profiling Modules

Profiling modules group similar network connections and search for dominant behaviors using clustering algorithms. Examples of profiling are categorized in Table 1.6. We address profiling techniques in Chapter 7.

Table 1.5 Examples of Data Mining for Scan Detection

Technique Used	*Granularity*	*Levels*	*References*
E.1 Statistical methods	Batch	Both	Staniford et al. (2002a,b)
E.2 Rule-based	Batch	Both	Staniford-Chen et al. (1996)
E.3 Threshold random walk	Continues	Host	Jung et al. (2004)
E.4 Expert knowledge—rule based	Batch	Network	Simon et al. (2006)
E.5 Associative memory	Continuous	Network	Muelder et al. (2007)

Table 1.6 Examples of Data Mining for Profiling

Technique Used	*Input Data Format*	*Levels*	*References*
F.1 Association rules	Set of network flow, offline	Network	Apiletti et al. (2008)
F.2 Shared nearest neighbor clustering (SNN)	Set of network flow, offline	Network	Ertöz et al. (2003), Chandola et al. (2006)
F.3 EM-based clustering	Set of network flow, offline	Network	Patcha and Park (2007)
F.4 Subspace	Set of network flow, offline	Network	Lakhina et al. (2004), Erman et al. (2006)
F.5 Information theoretic	Set of network flow, offline	Network	Xu et al. (2008)

1.5 Summary

In this chapter, we have introduced what we believe to be the most important components of cybersecurity, data mining, and machine learning. We provided an overview of types of cyber attacks and cybersecurity solutions and explained that cyber attacks compromise cyberinfrastructures in three ways: They help cyber criminals steal information, impair componential function, and disable services. We have briefly defined cybersecurity defense strategies, which consist of proactive and reactive solutions.

We highlighted proactive PPDM, and the reactive misuse detection, anomaly detection, and hybrid detection techniques. PPDM is rising in popularity as operative computation and data sharing in cyber space creates more concerns about privacy leaks, and misuse detection, anomaly detection, and hybrid detection techniques compose many IDSs. Misuse detection methods attempt to match test data with the profiled anomalous patterns, while anomalous detection solutions profile normal patterns to search for outliers. Hybrid detection systems combine misuse and anomalous detection techniques to improve the detection rate and reduce the false-alarm rate. In addition, we discuss two specific research areas in cybersecurity: scan detection and network profiling. Scan detection is used to detect the precursor of attacks, such that its use can lead to the earlier deterrence of attacks or defenses. Profiling networks facilitate the administration and monitoring of cybersecurity through extraction, aggregation, and visualization tools.

1.6 Further Reading

Throughout this book, we assume that the readers are familiar with cyberinfrastructures, with network intrusions, and with elementary probability theory, information theory, and linear algebra. Although we present a readable product for readers to solve cybersecurity problems using data-mining and machine-learning paradigms, we will provide further reading that we feel is related to our content to supplement that basic knowledge.

The resources in the areas of data mining and machine learning in cyber security are rich and rapidly growing. We provide a succinct list of the principal references for data mining, machine learning, cybersecurity, and privacy. We also list related books at the end of this chapter for readers to access the related material easily. In the later chapters of the book, we list readings that address the specific problems corresponding to the chapter topics. Our general reading list follows. If you are familiar with the material, you can skip to Chapter 2.

The key important forums on cybersecurity include the *ACM International Conference on Computer Security* (*S&P*), the *IEEE Symposium on Security and Privacy*, the *International Conference on Security and Management*, the *ACM Special Interest Group on Management of Data* (SIGMOD), the *National Computer Security Conference*, the *USENIX Security Symposium*, the *ISOC Network and Distributed System Security Symposium* (*NDSS*), the *International Conference on Security in Communication Networks*, the *Annual Computer Security Applications Conference*, the *International Symposium on Recent Advances in Intrusion Detection*, the *National Information Security Conference*, and the *Computer Security Foundations Workshop*.

The most important data-mining conferences include *ACM Knowledge Discovery and Data Mining*, *ACM Special Interest Group on Management of Data*, *Very Large Data Bases*, *IEEE International Conference on Data Mining*, *ACM Special Interest Group on Information Retrieval*, *IEEE International Conference on Data Engineering*, *International Conference on Database Theory*, and *Extending Database Technology*.

The most important machine-learning conferences include *American Association for AI National Conference* (*AAAI*), (*NIPS*), (*IJCAI*), *CVPR*, and *ICML*.

The most important journals on cybersecurity include *ACM Transactions on Information and System Security*, *IEEE Transactions on Dependable and Secure Computing*, *IEEE Transactions on Information Forensics and Security*, *Journal of Computer Security*, and the *International Journal of Information Security*.

The most important journals on data mining and machine learning include *IEEE Transactions on Pattern Analysis and Machine Learning*, *IEEE Transactions on Systems, Man and Cybernetics*, *IEEE Transactions on Software Engineering*, *IEEE/ACM Transactions on Networking*, *IEEE Transactions on Computers*, *IEEE Transactions on Knowledge and Data Engineering*, *Machine Learning Journal*, *Journal of Machine Learning Research*, *Neural Computation*, *Pattern Recognition*, and *Pattern Recognition Letters*.

We list a number of books that contain complementary knowledge in data mining, machine learning, and cybersecurity. These books provide readable and explanatory materials for readers to access.

Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach* (3rd edition), Prentice Hall, Upper Saddle River, NJ, 2009.

Stephen Northcutt and Judy Novak, *Network Intrusion Detection* (3rd edition), New Riders, Indianapolis, IN, 2003.

Daniel Barbará and Sushil Jajodia, *Applications of Data Mining in Computer Security*, Kluwer, Norwell, MA, 2002.

Tom Mitchell, *Machine Learning*, McGraw Hill, New York, 1997.

Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification* (2nd edition), Wiley, New York, 2001.

Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer, Heidelberg, 2006.

Jiawei Han and Micheline Kamber, *Data Mining Concepts and Techniques*, Morgan Kaufmann, San Francisco, CA, 2001.

David J. Hand, Heikki Mannila, and Padhraic Smyth, *Principles of Data Mining*, MIT Press, Cambridge, MA, 2001.

David J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, Cambridge, U.K., 2003.

Jaideep Vaidya, Christopher W. Clifton, and Yu Michael Zhu, *Privacy Preserving Data Mining*, Springer, New York, 2006.

References

Abraham, A., C. Grosan, and C. Martin-Vide. Evolutionary design of intrusion detection programs. In: *International Journal of Networks Security* 4 (3) (2007a): 328–339.

Abraham, A., R. Jain, J. Thomas, and S.Y. Han. DSCIDS: Distributed softcomputing intrusion detection system. *Journal of Network and Computer Applications* 30 (1) (2007b): 381–398.

Agrawal, R. and R. Srikant. Privacy-preserving data mining. In: *Proceedings of the ACM SIGMOD Conference on Management of Data*, Dallas, TX, 2000, pp. 439–450.

Apiletti, D., E. Baralis, T. Cerquitelli, and V. D'Elia. Characterizing network traffic by means of the NetMine framework. *Computer Networks* 53 (6) (2008): 774–789.

Axelsson, S. *Intrusion Detection Systems: A Survey and Taxonomy*. Göteborg, Sweden: Department of Computer Engineering, Chalmers University, 2000.

Barak, B., A. Herzberg, D. Naor, and E. Shai. The proactive security toolkit and applications. In: *Proceedings of the 6th ACM Conference on Computer and Communications Security*, Singapore, 1999, pp. 18–27.

Barni, M., C. Orlandi, and A. Piva. A privacy-preserving protocol for neural-network-based computation. In: *Proceedings of the 8th Workshop on Multimedia and Security*, Geneva, Switzerland, 2006, pp. 146–151.

Canetti, R., R. Gennaro, A. Herzberg, and D. Naor. Proactive security: Long-term protection against break-ins. *CryptoBytes* 3 (1997): 1–8.

Cannady, J. Artificial neural networks for misuse detection. In: *Proceedings of the 1998 National Information Systems Security Conference* (*NISSC'98*), Arlington, VA, 1998, pp. 443–456.

Chandola, V., E. Banerjee et al. Data mining for cyber security. In: *Data Warehousing and Data Mining Techniques for Computer Security*, edited by A. Singhal. Springer, New York, 2006.

Chebrolu, S., A. Abraham, and J.P. Thomas. Feature deduction and ensemble design of intrusion detection systems. *Computers & Security* 24 (2005): 1–13.

Chen, W.H., S.H. Hsu, and H.P. Shen. Application of SVM and ANN for intrusion detection. *Computers & Operations Research* 32 (2005): 2617–2634.

Cuppens, F. and A. Miège. Alert correlation in a cooperative intrusion detection framework. *IEEE Symposium on Research in Security and Privacy*, Oakland, CA, 2002.

Dain, O. and R. Cunningham. Building scenarios from a heterogeneous alert stream. In: *Proceedings of the 2001 IEEE Workshop on Information Assurance and Security*, West Point, NY, 2001a, pp. 231–235.

Dain, O. and R. Cunningham. Fusing a heterogeneous alert stream into scenarios. In: *Proceedings of the 2001, ACM Workshop on Data Mining for Security Applications*, Philadelphia, PA, 2001b, pp. 1–13.

Debar, H., M. Dacier, and A. Wespi. Toward taxonomy of intrusion detection systems. *Computer Networks* 31 (1999): 805–822.

Denning, D. An intrusion-detection model. *IEEE Transactions on Software Engineering* 13 (2) (1987): 118–131.

Du, W. and Z. Zhan. Building decision tree classifier on private data. In: *Proceedings of the IEEE ICDM Workshop on Privacy, Security and Data Mining*, Maebashi City, Japan, 2002.

Du, W., Y.S. Han, and S. Chen. Privacy-preserving multivariate statistical analysis: Linear regression and classification. In: *Proceedings of SIAM International Conference on Data Mining* (*SDM*), Lake Buena Vista, FL, 2004.

Endler, D. Intrusion detection: Applying machine learning to solaris audit data. In: *Proceedings of the 1998 Annual Computer Security Applications Conference* (*ACSAC*), Los Alamitos, CA, 1998, pp. 268–279.

Erman, J., M. Arlitt, and A. Mahanti. Traffic classification using clustering algorithms. In: *Proceedings of the 2006 ACM SIGCOMM Workshop on Mining Network Data*, Pisa, Italy, 2006.

Ertöz, L., M. Steinbach, and V. Kumar. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In: *Proceedings of the Third SIAM International Conference on Data Mining*, San Francisco, CA, 2003, pp. 47–58.

Evfimievski, A., R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In: *Proceedings of the 8th ACM SIGKDDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Canada, 2002.

Feinstein, L., D. Schnackenberg, R. Balupari, and D. Kindred. Statistical approaches to DdoS attack detection and response. In: *Proceedings of DARPA Information Survivability Conference and Exposition*, Washington, DC, 2003, pp. 303–314.

Ghosh, A.K., J. Wanken, and F. Charron. Detecting anomalous and unknown intrusions against programs. In: *Proceedings of the 1998 Annual Computer Security Applications Conference* (*ACSAC*), Scottsdale, AZ, 1998.

Ghosh, A.K., and A. Schwartzbard. A study in using neural networks for anomaly and misuse detection. In: *Proceedings of the 8th USENIX Security Symposium*, Washington, DC, 1999, pp. 141–152.

Ghosh, A.K., A. Schwartzbard, and M. Schatz. Learning program behavior profiles for intrusion detection USENIX Association. In: *Proceedings of the 1st USENIX Workshop on Intrusion Detection and Network Monitoring*, Santa Clara, CA, 1999.

Hodge, V.J. and J. Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review* 22 (2) (2004): 85–126.

Homeland Security Council. *National Strategy for Homeland Security*. 2007, p. 36, http://www.dhs.gov/xlibrary/assets/nat_strat_homeland-security_2007.pdf

Hu, W.J., Y.H. Liao, and V.R. Vemuri. Robust support vector machines for anomaly detection in computer security. In: *Proceedings of the International Conference on Machine Learning*, 2003, pp. 282–289.

Jung, J., V. Paxson, A.W. Berger, and H. Balakrishnan. Fast portscan detection using sequential hypothesis testing. In: *IEEE Symposium on Security and Privacy*, Oakland, CA, 2004.

Kantarcioglu, M. and C. Clifton. Privately computing a distributed k-nn classifier. In: *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Pisa, Italy, 2004, pp. 279–290.

Kruegel, C. and T. Toth. Using decision trees to improve signature-based intrusion detection. In: *Proceedings of the 6th International Workshop on the Recent Advances in Intrusion Detection*, West Lafayette, IN, 2003, pp. 173–191.

Lakhina, A., M. Crovella, and C. Diot. Characterization of network-wide anomalies in traffic flows. In: *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement*, Taormina, Sicily, Italy, 2004, pp. 201–206.

Lakhina, A., M. Crovella, and C. Diot. Mining anomalies using traffic feature distributions. In: *Proceedings of the 2005 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, Philadelphia, PA, 2005.

Lee, W. and W. Fan. Mining system audit data: Opportunities and challenges. *SIGMOD Record* 30 (4) (2001): 33–44.

Lee, W. and S.J. Stolfo. Data mining approaches for intrusion detection. In: *Proceedings of the 7th USENIX Security Symposium*, San Antonio, TX, 1998.

Lee, W. and S.J. Stolfo. A framework for constructing features and models for intrusion detection systems. *ACM Transactions on Information and System Security* (*TISSEC*) 2 (4) (2000): 227–261.

Lee, W., S.J. Stolfo, and K.W. Mok. A data mining framework for building intrusion detection models. In: *Proceedings of the IEEE Symposium on Security and Privacy*, 1999, pp. 120–132.

Leung, K. and C. Leckie. Unsupervised anomaly detection in network intrusion detection using clusters. In: *Proceedings of the Twenty-Eighth Australasian Conference on Computer Science*, 2005, pp. 333–342.

Li, X., F. Bian, M. Crovella, C. Diot, R. Govindan, G. Iannaccone, and A. Lakhina. Detection and identification of network anomalies using sketch subspaces. In: *Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement*, 2006, pp. 147–152.

Liao, Y.H. and V.R. Vemuri. Use of K-nearest neighbor classifier for intrusion detection. *Computers & Security* 21 (5) (2002): 439–448.

Liu, Z., G. Florez, and S.M. Bridges. A comparison of input representations in neural networks: A case study in intrusion detection. In: *Proceedings of the 2002 International Joint Conference on Neural Networks*, Honolulu, HI, 2002.

Mahoney, M.V. and P.K. Chan. Learning nonstationary models of normal network traffic for detecting novel attacks. In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Canada, 2002, pp. 376–386.

Mahoney, M.V. and P.K. Chan. Learning rules for anomaly detection of hostile network traffic. In: *Proceedings of the 3rd International Conference on Data Mining*, Melbourne, FL, 2003, pp. 601–603.

Meza, J., S. Campbell, and D. Bailey. *Mathematical and Statistical Opportunities in Cybersecurity*, Paper LBNL-1667E, Lawrence Berkeley National Laboratory, Berkeley, CA, 2009.

Muelder, C., L. Chen, R. Thomason, K.L. Ma, and T. Bartoletti. Intelligent classification and visualization of network scans. In: *Proceedings of the Workshop on Visualization for Computer Security*, Sacramento, CA, 2007.

Mukkamala, S. and A.H. Sung. A comparative study of techniques for intrusion detection. In: *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence*, 2003, pp. 570–577.

Ning, P., D. Xu, C. Healey, and R.S. Amant. Building attack scenarios through integration of complementary alert correlation method. In: *Proceedings of the 11th Annual Network and Distributed System Security Symposium*, San Diego, CA, 2004.

Patcha, A. and J.M. Park. An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer Networks* 51 (12) (2007): 3448–3470.

Porras, P.A. and P.G. Neumann. EMERALD: Event monitoring enabling responses to anomalous live disturbances. In: *Proceedings of the Nineteenth Computer Security*, Baltimore, MD, 1997, pp. 353–365.

Portnoy, L., E. Eskin, and S. Stolfo. Intrusion detection with unlabeled data using clustering. In: *Proceedings of ACM CSS Workshop on Data Mining Applied to Security* (*DMSA*), Philadelphia, PA, 2001.

Ringberg, H., A. Soule, J. Rexford, and C. Diot. Sensitivity of PCA for traffic anomaly detection. *ACM SIGMETRICS Performance Evaluation Review* 35 (1) (2007): 109–120.

Schultz, M.G., E. Eskin, E. Zadok, and S.J. Stolfo. Data mining methods for detection of new malicious executables. In: *DARPA Information Survivability Conference and Exposition* (*DISCEX*), Anaheim, CA, 2001.

Siddiqui, M., M.C. Wang, and J. Lee. A survey of data mining techniques for malware detection using file features. In: *Proceedings of the 46th Annual Southeast Regional Conference*, Auburn, Canada, 2008.

Simon, G., H. Xiong, E. Eilertson, and V. Kumar. Scan detection: A data mining approach. In: *Proceedings of the Sixth SIAM International Conference on Data Mining* (*SDM*), Bethesda, MD, 2006, pp. 118–129.

Smaha, S.E. Haystack: An intrusion detection system. In: *IEEE Fourth Aerospace Computer Security Applications Conference*, Orlando, FL, 1988, pp. 37–44.

Soule, A., K. Salamatian, and N. Taft. Combining filtering and statistical methods for anomaly detection. In: *Proceedings of the Fifth ACM SIGCOMM Conference on Internet Measurement*, Berkeley, CA, 2005.

Srinivas, M., S. Andrew, A. Ajith, and R. Vitorino. Intrusion detection systems using adaptive regression splines. In: *The Sixth International Conference on Enterprise Information Systems*, Porto, Portugal, 2004.

Staniford, S., J.A. Hoagland, and J.M. McAlerney. Practical automated detection of stealthy portscans. *Journal of Computer Security* 10 (2002a): 105–136.

Staniford, S., J.A. Hoagland, and J.M. McAlerney. Practical automated detection of stealthy portscans. In: *Proceedings of the 7th ACM Conference on Computer and Communications Security*, Athens, Greece, 2002b.

Staniford-Chen, S., S. Cheung, R. Crawford, M. Dilger, J. Frank, J. Hoagland, K. Levitt, C. Wee, R. Yip, and D. Zerkle. GrIDS: A graph-based intrusion detection system for large networks. In: *The 19th National Information Systems Security Conference*, Baltimore, MD, 1996.

Stolfo, S.J., W. Lee, P.K. Chan, W. Fan, and E. Eskin. Data mining-based intrusion detectors: An overview of the Columbia IDS project. *ACM SIGMOD Record* 30 (4) (2001): 5–14.

Su, M., G. Yu, and C. Lin. A real-time network intrusion detection system for large-scale attacks based on an incremental mining approach. *Computers and Security* 28 (5) (2009): 301–309.

Vaidya, J. and C. Clifton. Privacy-preserving k-means clustering over vertically partitioned data. In: *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington, DC, 2003.

Verykios, V.S., E. Bertino, I.N Fovino, L.P. Provenza, Y. Saygin, and Y. Theodoridis. State-of-the-art in privacy preserving data mining. *ACM SIGMOD Record* 33 (1) (2004): 50–57.

Warrender, C., S. Forrrest, and B. Pearlmutter. Detecting intrusions using system calls: Alternative data models. In: *IEEE Symposium on Security and Privacy*, Oakland, CA, 1999, pp. 133–145.

Wright, R. and Z. Yang. Privacy-preserving Bayesian network structure computation on distributed heterogeneous data. In: *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, WA, 2004.

Xu, K., X.L. Zhang, and S. Bhattachayya. Internet traffic behavior profiling for network security monitoring. *IEEE/ACM Transactions on Networking* (*TON*) 16 (6) (2008): 1241–1252.

Yamanishi, K. and J.I. Takeuchi. Discovering outlier filtering rules from unlabeled data: Combining a supervised learner with an unsupervised learner. In: *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, 2001, pp. 389–394.

Yamanishi, K., J.I. Takeuchi, G. Williams, and P. Milne. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. In: *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery*, Boston, MA, 2000.

Ye, N., X.Y. Li, Q. Chen, S.M. Emran, and M.M Xu. Probabilistic techniques for intrusion detection based on computer audit data. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans* 31 (4) (2001): 266–274.

Ye, N., S.M. Emran, Q. Chen, and S. Vilbert. Multivariate statistical analysis of audit trails for host-based intrusion detection. *IEEE Transactions on Computers* 51 (2002): 810–820.

Yu, H., X. Jiang, and J. Vaidya. Privacy-preserving SVM using nonlinear kernels on horizontally partitioned data. In: *Proceedings of the 2006 ACM Symposium on Applied Computing*, Dijon, France, 2006.

Zhang, J. and M. Zulkernine. A hybrid network intrusion detection technique using random forests. In: *Proceedings of the First International Conference on Availability, Reliability and Security*, 2006a, pp. 262–269.

Zhang, J. and M. Zulkernine. Anomaly based network intrusion detection with unsupervised outlier detection. In: *IEEE International Conference on Communications*, Istanbul, Turkey, 2006b.

References

1 Chapter 1: Introduction

B.1 Rule-based signature analysis Frequency of system calls, off line Host Lee et al. (1999)

B.2 ANN TCP/IP data, offline Host Ghosh and Schwartzbard (1999), Cannady (1998)

B.3 Fuzzy association rules Frequency of system calls, online Host Abraham et al. (2007b), Su et al. (2009)

B.4 SVM TCP/IP data, offline Network Mukkamala and Sung (2003)

B.5 Linear genetic programs (LGP) TCP/IP data, offline Network Mukkamala and Sung (2003), Abraham et al. (2007a,b), Srinivas et al. (2004)

B.6 Classification and regression trees Frequency of system calls, offline Host Chebrolu et al. (2005)

B.7 Decision tree TCP/IP data, online Network Kruegel and Toth (2003)

B.8 BN Frequency of system calls, offline Host Chebrolu et al. (2005)

B.9 Statistical method Executables, offline Host Schultz et al. (2001)

12 ▪

Table 1.3 Examples of Data Mining and Machine Learning

for Anomaly Detection

C.1 Statistical methods Sequences of system calls, offline Host Ye et al. (2001), Feinstein et al. (2003), Smaha (1988), Ye et al. (2002)

C.2 Statistical methods TCP/IP data, online Network Yamanishi and Takeuchi (2001), Yamanishi et al. (2000), Mahoney and Chan (2002, 2003), Soule et al. (2005)

C.3 Unsupervised clustering algorithm TCP/IP data, offline Network Portnoy et al. (2001), Leung and Leckie (2005), Warrender et al. (1999), Zhang and Zulkernine (2006a,b)

C.4 Subspace TCP/IP data offline Network Li et al. (2006)

C.5 Information theoretic TCP/IP, online Network Lakhina et al. (2005)

C.6 Association rules Frequency of system calls, online Host Lee and Stolfo (1998), Abraham et al. (2007a,b), Su et al. (2009), Lee et al. (1999)

C.7 Kalman filter TCP/IP data, online Network Soule et al. (2005)

C.8 Hidden Markov model (HMM) Sequences of system calls, offline Host Warrender et al. (1999)

C.9 ANN Sequences of system calls, offline Host Ghosh et al. (1998, 1999), Liu et al. (2002)

C.10 Principal component analysis (PCA) TCP/IP data, online Network Lakhina et al. (2004), Ringberg et al. (2007)

C.11 KNN Frequency of system calls, offline Host Liao and Vemuri (2002)

C.12 SVM TCP/IP data, offline Network Hu et al. (2003), Chen et al. (2005) ▪

1.4.2.3 Hybrid Detection

Most current IDSs employ either misuse detection techniques or anomaly detection

techniques. Both of these methods have drawbacks: misuse detection techniques

lack the ability to detect unknown intrusions; anomaly detection techniques usu

ally produce a high percentage of false alarms. To improve the techniques of IDSs,

researchers have proposed hybrid detection techniques to combine anomaly and

misuse detection techniques in IDSs. Examples for hybrid detection techniques are

listed in Table 1.4. We address hybrid detection techniques in Chapter 5.

1.4.2.4 Scan Detection

Scan detection generates alerts when attackers scan services or computer compo

nents in network systems before launching attacks. A scan detector identi�es the

precursor of an attack on a network, e.g., destination IPs and the source IPs of

Internet connections. Although many scan detection techniques have been pro

posed and declared to be able to detect the precursors of cyber attacks, the high

false-positive rate or the low scan detection rate limits the application of these solu

tions in practice. Some examples of scan detection techniques are categorized in

Table 1.5. We address scan and scan detection techniques in Chapter 6.

1.4.2.5 Profiling Modules

Pro�ling modules group similar network connections and search for dominant

behaviors using clustering algorithms. Examples of pro�ling are categorized in

Table 1.6. We address pro�ling techniques in Chapter 7.

Table 1.4 Examples of Data Mining for Hybrid Intrusion Detection

D.1 Correlation TCP/IP data, online Network Ning et al. (2004), Cuppens and Miège (2002), Dain and Cunningham (2001a,b)

D.2 Statistical methods Sequences of system calls, offline Host Endler (1998)

D.3 ANN Sequences of system calls, offline Host Endler (1998)

D.4 Association rules Frequency of system calls, online Host Lee and Stolfo (2000)

D.5 ANN TCP/IP data, online Network Ghosh et al. (1999)

D.6 Random forest TCP/IP data, online Network Zhang and Zulkernine (2006a,b)

1.5 Summary

In this chapter, we have introduced what we believe to be the most important

components of cybersecurity, data mining, and machine learning. We provided

an overview of types of cyber attacks and cybersecurity solutions and explained

that cyber attacks compromise cyberinfrastructures in three ways: �ey help cyber

criminals steal information, impair componential function, and disable services.

We have brie¤y de�ned cybersecurity defense strategies, which consist of proactive

and reactive solutions. We highlighted proactive PPDM, and the reactive misuse detection, anom

aly detection, and hybrid detection techniques. PPDM is rising in popularity as

Table 1.5 Examples of Data Mining for Scan Detection

E.1 Statistical methods Batch Both Staniford et al. (2002a,b)

E.2 Rule-based Batch Both Staniford-Chen et al. (1996)

E.3 Threshold random walk Continues Host Jung et al. (2004)

E.4 Expert knowledge—rule based Batch Network Simon et al. (2006)

E.5 Associative memory Continuous Network Muelder et al. (2007)

Table 1.6 Examples of Data Mining for Profiling

F.1 Association rules Set of network flow, offline Network Apiletti et al. (2008)

F.2 Shared nearest neighbor clustering (SNN) Set of network flow, offline Network Ertöz et al. (2003), Chandola et al. (2006)

F.3 EM-based clustering Set of network flow, offline Network Patcha and Park (2007)

F.4 Subspace Set of network flow, offline Network Lakhina et al. (2004), Erman et al. (2006)

F.5 Information theoretic Set of network flow, offline Network Xu et al. (2008) ▪

operative computation and data sharing in cyber space creates more concerns about

privacy leaks, and misuse detection, anomaly detection, and hybrid detection tech

niques compose many IDSs. Misuse detection methods attempt to match test data

with the pro�led anomalous patterns, while anomalous detection solutions pro�le

normal patterns to search for outliers. Hybrid detection systems combine misuse

and anomalous detection techniques to improve the detection rate and reduce the

false-alarm rate. In addition, we discuss two speci�c research areas in cybersecurity:

scan detection and network pro�ling. Scan detection is used to detect the precursor

of attacks, such that its use can lead to the earlier deterrence of attacks or defenses.

Pro�ling networks facilitate the administration and monitoring of cybersecurity through extraction, aggregation, and visualization tools.

1.6 Further Reading

�roughout this book, we assume that the readers are familiar with cyberinfra

structures, with network intrusions, and with elementary probability theory, infor

mation theory, and linear algebra. Although we present a readable product for

readers to solve cybersecurity problems using data-mining and machine-learning

paradigms, we will provide further reading that we feel is related to our content to

supplement that basic knowledge. �e resources in the areas of data mining and machine learning in cyber secu

rity are rich and rapidly growing. We provide a succinct list of the principal refer

ences for data mining, machine learning, cybersecurity, and privacy. We also list

related books at the end of this chapter for readers to access the related material

easily. In the later chapters of the book, we list readings that address the speci�c

problems corresponding to the chapter topics. Our general reading list follows. If

you are familiar with the material, you can skip to Chapter 2. �e key important forums on cybersecurity include the ACM International

Conference on Computer Security (S&P), the IEEE Symposium on Security and

Privacy, the International Conference on Security and Management, the ACM Special

Interest Group on Management of Data (SIGMOD), the National Computer Security

Conference, the USENIX Security Symposium, the ISOC Network and Distributed System Security Symposium (NDSS), the International Conference on Security in Communication Networks, the Annual Computer Security Applications Conference, the International Symposium on Recent Advances in Intrusion Detection, the National Information Security Conference, and the Computer Security Foundations Workshop. The most important data-mining conferences include ACM Knowledge Discovery and Data Mining, ACM Special Interest Group on Management of Data, Very Large Data Bases, IEEE International Conference on Data Mining, ACM Special Interest Group on Information Retrieval, IEEE International Conference on Data Engineering, International Conference on Database Theory, and Extending Database Technology. The most important machine-learning conferences include American Association for AI National Conference (AAAI), (NIPS), (IJCAI), CVPR, and ICML.

16 . The most important journals on cybersecurity include ACM Transactions on Information and System Security, IEEE Transactions on Dependable and Secure Computing, IEEE Transactions on Information Forensics and Security, Journal of Computer Security, and the International Journal of Information Security. The most important journals on data mining and machine learning include IEEE Transactions on Pattern Analysis and Machine Learning, IEEE Transactions on Systems, Man and Cybernetics, IEEE Transactions on

Software Engineering, IEEE/ACM Transactions on Networking, IEEE Transactions on Computers, IEEE Transactions on Knowledge and Data Engineering, Machine Learning Journal, Journal of Machine Learning Research, Neural Computation, Pattern Recognition, and Pattern Recognition Letters. We list a number of books that contain complementary knowledge in data mining, machine learning, and cybersecurity. These books provide readable and explanatory materials for readers to access.

Stuart J. Russell and Peter Norvig, Artificial Intelligence: A Modern Approach (3rd edition), Prentice Hall, Upper Saddle River, NJ, 2009.

Stephen Northcutt and Judy Novak, Network Intrusion Detection (3rd edition), New Riders, Indianapolis, IN, 2003.

Daniel Barbará and Sushil Jajodia, Applications of Data Mining in Computer Security, Kluwer, Norwell, MA, 2002.

Tom Mitchell, Machine Learning, McGraw Hill, New York, 1997.

Richard O. Duda, Peter E. Hart, and David G. Stork, Pattern Classification (2nd edition), Wiley, New York, 2001.

Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer, Heidelberg, 2006.

Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann, San Francisco, CA, 2001.

David J. Hand, Heikki Mannila, and Padhraic Smyth, Principles of Data Mining,

MIT Press, Cambridge, MA, 2001.

David J. C. MacKay, Information Theory, Inference, and Learning Algorithms,

Cambridge University Press, Cambridge, U.K., 2003.

Jaideep Vaidya, Christopher W. Clifton, and Yu Michael Zhu, Privacy Preserving

Data Mining, Springer, New York, 2006.

Abraham, A., C. Grosan, and C. Martin-Vide. Evolutionary design of intrusion detection programs. In: International Journal of Networks Security 4 (3) (2007a): 328–339.

Abraham, A., R. Jain, J. Thomas, and S.Y. Han. DSCIDS: Distributed softcomputing intrusion detection system. Journal of Network and Computer Applications 30 (1) (2007b): 381–398. ▪

Agrawal, R. and R. Srikant. Privacy-preserving data mining. In: Proceedings of the ACM SIGMOD Conference on Management of Data, Dallas, TX, 2000, pp. 439–450.

Apiletti, D., E. Baralis, T. Cerquitelli, and V. D'Elia. Characterizing network traffic by means of the NetMine framework. Computer Networks 53 (6) (2008): 774–789.

Axelsson, S. Intrusion Detection Systems: A Survey and Taxonomy. Göteborg, Sweden: Department of Computer Engineering, Chalmers University, 2000.

Barak, B., A. Herzberg, D. Naor, and E. Shai. The proactive security toolkit and applications. In: Proceedings of the 6th ACM Conference on Computer and Communications Security, Singapore, 1999, pp. 18–27.

Barni, M., C. Orlandi, and A. Piva. A privacy-preserving protocol for neural-network-based computation. In: Proceedings of the 8th Workshop on Multimedia and Security, Geneva, Switzerland, 2006, pp. 146–151.

Canetti, R., R. Gennaro, A. Herzberg, and D. Naor. Proactive security: Long-term protection against break-ins. CryptoBytes 3 (1997): 1–8.

Cannady, J. Artificial neural networks for misuse detection. In: Proceedings of the 1998 National Information Systems Security Conference (NISSC ’98), Arlington, VA, 1998, pp. 443–456.

Chandola, V., E. Banerjee et al. Data mining for cyber security. In: Data Warehousing and Data Mining Techniques for Computer Security, edited by A. Singhal. Springer, New York, 2006.

Chebrolu, S., A. Abraham, and J.P. Thomas. Feature deduction and ensemble design of intrusion detection systems. Computers & Security 24 (2005): 1–13.

Chen, W.H., S.H. Hsu, and H.P. Shen. Application of SVM and ANN for intrusion detection. Computers & Operations Research 32 (2005): 2617–2634.

Cuppens, F. and A. Miège. Alert correlation in a cooperative intrusion detection framework. IEEE Symposium on Research in Security and Privacy, Oakland, CA, 2002.

Dain, O. and R. Cunningham. Building scenarios from a heterogeneous alert stream. In: Proceedings of the 2001 IEEE Workshop on Information Assurance and Security, West Point, NY, 2001a, pp. 231–235.

Dain, O. and R. Cunningham. Fusing a heterogeneous alert stream into scenarios. In: Proceedings of the 2001, ACM Workshop on Data Mining for Security Applications, Philadelphia, PA, 2001b, pp. 1–13.

Debar, H., M. Dacier, and A. Wespi. Toward taxonomy of intrusion detection systems. Computer Networks 31 (1999): 805–822.

Denning, D. An intrusion-detection model. IEEE Transactions on Software Engineering 13 (2) (1987): 118–131.

Du, W. and Z. Zhan. Building decision tree classifier on private data. In: Proceedings of the IEEE ICDM Workshop on Privacy, Security and Data Mining, Maebashi City, Japan, 2002.

Du, W., Y.S. Han, and S. Chen. Privacy-preserving multivariate statistical analysis: Linear regression and classification. In: Proceedings of SIAM International Conference on Data Mining (SDM), Lake Buena Vista, FL, 2004.

Endler, D. Intrusion detection: Applying machine learning to solaris audit data. In: Proceedings of the 1998 Annual Computer Security Applications Conference (ACSAC), Los Alamitos, CA, 1998, pp. 268–279.

Erman, J., M. Arlitt, and A. Mahanti. Tra¢c classi▮cation using clustering algorithms. In: Proceedings of the 2006 ACM SIGCOMM Workshop on Mining Network Data, Pisa, Italy, 2006.

Ertöz, L., M. Steinbach, and V. Kumar. Finding clusters of di®erent sizes, shapes, and densities in noisy, high dimensional data. In: Proceedings of the ▮ird SIAM International Conference on Data Mining, San Francisco, CA, 2003, pp. 47–58.

Ev▮mievski, A., R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In: Proceedings of the 8th ACM SIGKDDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Canada, 2002.

Feinstein, L., D. Schnackenberg, R. Balupari, and D. Kindred. Statistical approaches to DdoS attack detection and response. In: Proceedings of DARPA Information Survivability Conference and Exposition, Washington, DC, 2003, pp. 303–314.

Ghosh, A.K., J. Wanken, and F. Charron. Detecting anomalous and unknown intrusions against programs. In: Proceedings of the 1998 Annual Computer Security Applications Conference (ACSAC), Scottsdale, AZ, 1998.

Ghosh, A.K., and A. Schwartzbard. A study in using neural networks for anomaly and misuse detection. In: Proceedings of the 8th USENIX Security Symposium, Washington, DC, 1999, pp. 141–152.

Ghosh, A.K., A. Schwartzbard, and M. Schatz. Learning program behavior pro▮les for intrusion detection USENIX Association. In: Proceedings of the 1st USENIX Workshop on Intrusion Detection and Network Monitoring, Santa Clara, CA, 1999.

Hodge, V.J. and J. Austin. A survey of outlier detection methodologies. Arti▮cial Intelligence Review 22 (2) (2004): 85–126.

Homeland Security Council. National Strategy for Homeland Security. 2007, p. 36, http://

Hu, W.J., Y.H. Liao, and V.R. Vemuri. Robust support vector machines for anomaly detection in computer security. In: Proceedings of the International Conference on Machine Learning, 2003, pp. 282–289.

Jung, J., V. Paxson, A.W. Berger, and H. Balakrishnan. Fast portscan detection using sequential hypothesis testing. In: IEEE Symposium on Security and Privacy, Oakland, CA, 2004.

Kantarcioglu, M. and C. Clifton. Privately computing a distributed k-nn classiﬁer. In: Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases, Pisa, Italy, 2004, pp. 279–290.

Kruegel, C. and T. Toth. Using decision trees to improve signature-based intrusion detection. In: Proceedings of the 6th International Workshop on the Recent Advances in Intrusion Detection, West Lafayette, IN, 2003, pp. 173–191.

Lakhina, A., M. Crovella, and C. Diot. Characterization of network-wide anomalies in tra¢c ¤ows. In: Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement, Taormina, Sicily, Italy, 2004, pp. 201–206.

Lakhina, A., M. Crovella, and C. Diot. Mining anomalies using tra¢c feature distributions. In: Proceedings of the 2005 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, Philadelphia, PA, 2005.

Lee, W. and W. Fan. Mining system audit data: Opportunities and challenges. SIGMOD Record 30 (4) (2001): 33–44.

Lee, W. and S.J. Stolfo. Data mining approaches for intrusion detection. In: Proceedings of the 7th USENIX Security Symposium, San Antonio, TX, 1998.

Lee, W. and S.J. Stolfo. A framework for constructing features and models for intrusion detection systems. ACM Transactions on Information and System Security (TISSEC) 2 (4) (2000): 227–261.

Lee, W., S.J. Stolfo, and K.W. Mok. A data mining framework for building intrusion detection models. In: Proceedings of the IEEE Symposium on Security and Privacy, 1999, pp. 120-132.

Leung, K. and C. Leckie. Unsupervised anomaly detection in network intrusion detection using clusters. In: Proceedings of the Twenty-Eighth Australasian Conference on Computer Science, 2005, pp. 333–342.

Li, X., F. Bian, M. Crovella, C. Diot, R. Govindan, G. Iannaccone, and A. Lakhina. Detection and identification of network anomalies using sketch subspaces. In: Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement, 2006, pp. 147–152.

Liao, Y.H. and V.R. Vemuri. Use of K-nearest neighbor classifier for intrusion detection. Computers & Security 21 (5) (2002): 439–448.

Liu, Z., G. Florez, and S.M. Bridges. A comparison of input representations in neural networks: A case study in intrusion detection. In: Proceedings of the 2002 International Joint Conference on Neural Networks, Honolulu, HI, 2002.

Mahoney, M.V. and P.K. Chan. Learning nonstationary models of normal network traffic for detecting novel attacks. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Canada, 2002, pp. 376–386.

Mahoney, M.V. and P.K. Chan. Learning rules for anomaly detection of hostile network traffic. In: Proceedings of the 3rd International Conference on Data Mining, Melbourne, FL, 2003, pp. 601–603.

Meza, J., S. Campbell, and D. Bailey. Mathematical and Statistical Opportunities in Cybersecurity, Paper LBNL-1667E, Lawrence Berkeley National Laboratory, Berkeley, CA, 2009.

Muelder, C., L. Chen, R. Thomason, K.L. Ma, and T. Bartoletti. Intelligent classification and visualization of network scans. In: Proceedings of the Workshop on Visualization for Computer Security, Sacramento, CA, 2007.

Mukkamala, S. and A.H. Sung. A comparative study of techniques for intrusion detection. In: Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence, 2003, pp. 570–577.

Ning, P., D. Xu, C. Healey, and R.S. Amant. Building attack scenarios through integration of complementary alert correlation method. In: Proceedings of the 11th Annual Network and Distributed System Security Symposium, San Diego, CA, 2004.

Patcha, A. and J.M. Park. An overview of anomaly detection techniques: Existing solutions and latest technological trends. Computer Networks 51 (12) (2007): 3448–3470.

Porras, P.A. and P.G. Neumann. EMERALD: Event monitoring enabling responses to anomalous live disturbances. In: Proceedings of the Nineteenth Computer Security, Baltimore, MD, 1997, pp. 353–365.

Portnoy, L., E. Eskin, and S. Stolfo. Intrusion detection with unlabeled data using clustering. In: Proceedings of ACM CSS Workshop on Data Mining Applied to Security (DMSA), Philadelphia, PA, 2001.

Ringberg, H., A. Soule, J. Rexford, and C. Diot. Sensitivity of PCA for traffic anomaly detection. ACM SIGMETRICS Performance Evaluation Review 35 (1) (2007): 109–120.

Schultz, M.G., E. Eskin, E. Zadok, and S.J. Stolfo. Data mining methods for detection of new malicious executables. In: DARPA Information Survivability Conference and Exposition (DISCEX), Anaheim, CA, 2001.

Siddiqui, M., M.C. Wang, and J. Lee. A survey of data mining techniques for malware detection using file features. In: Proceedings of the 46th Annual Southeast Regional Conference, Auburn, Canada, 2008.

Simon, G., H. Xiong, E. Eilertson, and V. Kumar. Scan detection: A data mining approach. In: Proceedings of the Sixth SIAM International Conference on Data Mining (SDM), Bethesda, MD, 2006, pp. 118–129.

Smaha, S.E. Haystack: An intrusion detection system. In: IEEE Fourth Aerospace Computer Security Applications Conference, Orlando, FL, 1988, pp. 37–44.

Soule, A., K. Salamatian, and N. Taft. Combining filtering and statistical methods for anomaly detection. In: Proceedings of the Fifth ACM SIGCOMM Conference on Internet Measurement, Berkeley, CA, 2005.

Srinivas, M., S. Andrew, A. Ajith, and R. Vitorino. Intrusion detection systems using adaptive regression splines. In: The Sixth International Conference on Enterprise Information Systems, Porto, Portugal, 2004.

Staniford, S., J.A. Hoagland, and J.M. McAlerney. Practical automated detection of stealthy portscans. Journal of Computer Security 10 (2002a): 105–136.

Staniford, S., J.A. Hoagland, and J.M. McAlerney. Practical automated detection of stealthy portscans. In: Proceedings of the 7th ACM Conference on Computer and Communications Security, Athens, Greece, 2002b.

Staniford-Chen, S., S. Cheung, R. Crawford, M. Dilger, J. Frank, J. Hoagland, K. Levitt, C. Wee, R. Yip, and D. Zerkle. GrIDS: A graph-based intrusion detection system for large networks. In: The 19th National Information Systems Security Conference, Baltimore, MD, 1996.

Stolfo, S.J., W. Lee, P.K. Chan, W. Fan, and E. Eskin. Data mining-based intrusion detectors: An overview of the Columbia IDS project. ACM SIGMOD Record 30 (4) (2001): 5–14.

Su, M., G. Yu, and C. Lin. A real-time network intrusion detection system for large-scale attacks based on an incremental mining approach. Computers and Security 28 (5) (2009): 301–309.

Vaidya, J. and C. Clifton. Privacy-preserving k-means clustering over vertically partitioned data. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, 2003.

Verykios, V.S., E. Bertino, I.N Fovino, L.P. Provenza, Y. Saygin, and Y. Theodoridis. State-ofthe-art in privacy preserving data mining. ACM SIGMOD Record 33 (1) (2004): 50–57.

Warrender, C., S. Forrrest, and B. Pearlmutter. Detecting intrusions using system calls: Alternative data models. In: IEEE Symposium on Security and Privacy, Oakland, CA, 1999, pp. 133–145.

Wright, R. and Z. Yang. Privacy-preserving Bayesian network structure computation on distributed heterogeneous data. In: Proceedings of the tenth ACM SIGKDD International

Conference on Knowledge Discovery and Data Mining, Seattle, WA, 2004.

Xu, K., X.L. Zhang, and S. Bhattachayya. Internet tra¢c behavior pro􀀀ling for network security monitoring. IEEE/ACM Transactions on Networking (TON) 16 (6) (2008): 1241–1252.

Yamanishi, K. and J.I. Takeuchi. Discovering outlier 􀀀ltering rules from unlabeled data: Combining a supervised learner with an unsupervised learner. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, 2001, pp. 389–394.

Yamanishi, K., J.I. Takeuchi, G. Williams, and P. Milne. On-line unsupervised outlier detection using 􀀀nite mixtures with discounting learning algorithms. In: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery, Boston, MA, 2000. ▪

Ye, N., X.Y. Li, Q. Chen, S.M. Emran, and M.M Xu. Probabilistic techniques for intrusion detection based on computer audit data. IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans 31 (4) (2001): 266–274.

Ye, N., S.M. Emran, Q. Chen, and S. Vilbert. Multivariate statistical analysis of audit trails for host-based intrusion detection. IEEE Transactions on Computers 51 (2002): 810–820.

Yu, H., X. Jiang, and J. Vaidya. Privacy-preserving SVM using nonlinear kernels on horizontally partitioned data. In: Proceedings of the 2006 ACM Symposium on Applied Computing, Dijon, France, 2006.

Zhang, J. and M. Zulkernine. A hybrid network intrusion detection technique using random forests. In: Proceedings of the First International Conference on Availability, Reliability and Security, 2006a, pp. 262–269.

Zhang, J. and M. Zulkernine. Anomaly based network intrusion detection with unsupervised outlier detection. In: IEEE International Conference on Communications, Istanbul, Turkey, 2006b.

2 Chapter 2: Classical Machine-Learning Paradigms for Data Mining

Breiman, L. Bagging predictors. Machine Learning 24 (2) (1996): 123–140.

Breiman, L. Random forests. Machine Learning 45 (1) (2001): 5–32.

Chapelle, O., B. Schölkopf, and A. Zien, eds. Semi-Supervised Learning. Cambridge, MA: �e MIT Press, 2006.

Dietterich, T.G. Machine learning for sequential data: A review. In: Proceedings of the Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition, Windsor and Ontario, Canada, 2002, pp. 15–30.

Freund, Y. and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55 (1) (1997): 119–139.

Freund, Y. and R.E. Schapire. A short introduction to boosting (in Japanese, translation by Naoki Abe). Journal of Japanese Society for Arti�cial Intelligence 14 (5) (1999): 771–780.

He, H. and E.A. Garcia. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering 21(9) (2009): 1263–1284.

Jain, A.K., R.P.W. Duin, and J. Mao. Statistical pattern recognition: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (1) (2000): 4–37.

Kaelbling, L.P., M.L. Littman, and A.W. Moore. Reinforcement learning: A survey. Journal of Arti�cial Intelligence Research 4 (1996): 237–285.

Lakhina, A., M. Crovella, and C. Diot. Diagnosing network-wide tra¢c anomalies. In: Proceedings of the 2004 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, Portland, OR, Vol. 34, No. 4, 2004, pp. 219–230.

Parsons, L., E. Haque, and H. Liu. Subspace clustering for high dimensional data: A review. SIGKDD Exploration

Newsletter 6 (1) (2004): 90-105.

Rabiner, L.R. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77 (2) (1989): 257-286.

Roweis, S. and Z. Ghahramani. A unifying review of linear Gaussian models. Neural Computation 11 (2) (1999): 305-345.

3 Chapter 3: Supervised Learning for Misuse/Signature Detection

Abraham, A., R. Jain, J. Thomas, and S.Y. Han. DSCIDS: Distributed softcomputing intrusion detection system. Journal of Network and Computer Applications 30 (1) (2007a): 381–398.

Abraham, A., C. Grosan, and C. Martin-Vide. Evolutionary design of intrusion detection programs. International Journal of Network Security 4 (3) (2007b): 328–339.

Agrawal, R., T. Imielinski, and A. Swami. Mining Association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD Conference, Washington, DC, 1993.

Cannady, J. Artificial neural networks for misuse detection. In: Proceedings of the 1998 National Information Systems Security Conference (NISSC'98), Arlington, VA, October 5–8, 1998, pp. 443–456.

Chebrolu, S., A. Abraham, and J.P. Thomas. Feature deduction and ensemble design of intrusion detection systems. Computers & Security 24 (2005): 1–13.

Cooper, G.F. and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. Machine Learning 9 (1992): 309–347.

Kruegel, C. and T. Toth. Using decision trees to improve signature-based intrusion detection. In: Proceedings of the 6th International Workshop on the Recent Advances in Intrusion Detection, Pittsburgh, PA, 2003, pp. 173–191.

Lee, W.K., S.J. Stolfo, and K.W. Mok. Mining audit data to build intrusion detection models. In: Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, New York, August 1998.

Lee, W.K., S.J. Stolfo, and K.W. Mok. A data mining framework for building intrusion detection models. In: Proceedings of the IEEE Symposium on Security and Privacy, Oakland, CA, 1999, pp. 120–132.

Mukkamala, S., G. Janoski, and A.H. Sung. Intrusion detection using support vector machines. In: Proceedings of Advanced Simulation Technologies Conference, 2002, pp. 178–183.

Mukkamala, S. and A.H. Sung. A comparative study of techniques for intrusion detection. In: Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence, Sacramento, CA, 2003, pp. 570–577.

Pearl, J. and N. Wermuch. When can association graphs admit a causal interpretation? In: Preliminary Papers of the Fourth International Workshop on Artificial Intelligence and Statistics, Fort Lauderdale, FL, 1993, pp. 141–150.

Russell, S.J. and P. Norvig. Artificial Intelligence: A Modern Approach, 2nd edn. Upper Saddle River, NJ: Prentice Hall, 2003, ISBN 0-13-790395-2.

Schultz, M.G., E. Eskin, E. Zadok, and S.J. Stolfo. Data mining methods for detection of new malicious executables. In: DARPA Information Survivability Conference and Exposition (DISCEX II'01), Anaheim, CA, Vol. 1, 2001.

Verma, T. and J. Pearl. An algorithm for deciding if a set of observed independencies has a causal explanation. In: Proceedings 8th Conference on Uncertainty in AI, Stanford, CA, 1992, pp. 323–330.

4 Chapter 4: Machine Learning for Anomaly Detection

Agrawal, R., J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, Seattle, WA, 1998, pp. 94–105.

Apiletti, D., E. Baralis, T. Cerquitelli, and V. D'Elia. Characterizing network traffic by means of the NetMine framework. Computer Networks 53 (6) (2008): 774–789.

Cannady, J. Artificial neural networks for misuse detection. In: Proceedings of the 1998 National Information Systems Security Conference (NISSC'98), Arlington, VA, 1998, pp. 443–456.

Chandola, V., E. Banerjee et al. Data mining for cyber security. In: Data Warehousing and Data Mining Techniques For Computer Security, edited by A. Singhal, Springer, New York, 2006.

Chen, W.H., S.H. Hsu, and H.P. Shen. Application of SVM and ANN for intrusion detection. Computers & Operations Research 32 (10) (2005): 2617–2634.

Eskin, E. Anomaly detection over noisy data using learned probability distributions. In: Proceedings of the International Conference on Machine Learning (ICML), Palo Alto, CA, 2000.

Eskin, E., A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo. A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. In: Applications of Data Mining in Computer Security, edited by S. Jajodia and D. Barbara. Dordrecht: Kluwer, 2002, Chap. 4.

Feinstein, L., D. Schnackenberg, R. Balupari, and D. Kindred. Statistical approaches to DDoS attack detection and response. In: Proceedings of DARPA Information Survivability Conference and Exposition, Washington, DC, 2003, pp. 303–314.

Ghosh, A.K., J. Wanken, and F. Charron. Detecting anomalous and unknown intrusions against programs. In: Proceedings of the 1998 Annual Computer Security Applications Conference (ACSAC), Phoenix, AZ, 1998.

Ghosh, A.K., A. Schwartzbard, and M. Schatz. Learning program behavior pro■les for intrusion detection USENIX Association. In: Proceedings of the 1st USENIX Workshop on Intrusion Detection and Network Monitoring, Santa Clara, CA, 1999.

Gong, F. Deciphering detection techniques: Part II. Anomaly-based intrusion detection. white paper, Mcafee Network Security Technologies Group, 2003.

Luis MartinGarcia, Tcpdump/Libpcap public repository, http://www.tcpdump.org/, Accessed on January 31, 2011.

Hu, W.J., Y.H. Liao, and V.R. Vemuri. Robust support vector machines for anomaly detection in computer security. In: Proceedings of the International Conference on Machine Learning, Las Vegas, NV, 2003, pp. 282–289.

Jackson, J.E. and G.S. Mudholkar. Control procedures for residuals associated with principal component analysis. Technometrics 21 (3) (1979): 341–349.

Jiang, S., X. Song, H. Wang, J. Han, and Q. Li. A clustering-based method for unsupervised intrusion detections. Pattern Recognition Letters 27 (7) (2006): 802–810.

Lakhina, A., M. Crovella, and C. Diot. Characterization of network-wide anomalies in tra¢c ¤ows. In: Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement, Taormina, Sicily, Italy, 2004a, pp. 201–206.

Lakhina, A., M. Crovella, and C. Diot. Diagnosing network-wide tra¢c anomalies. In: Proceedings of the 2004 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, Portland, OR, 2004b, pp. 219–230.

Lakhina, A., M. Crovella, and C. Diot. Mining anomalies using tra¢c feature distributions. In: Proceedings of the 2005 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, Philadelphia, PA, 2005.

Lee, W. and S.J. Stolfo. Data mining approaches for intrusion detection. In: Proceedings of the 7th USENIX Security Symposium, San Antonio, TX, 1998.

Lee, W. and Xiang, D. Information-theoretic measures for anomaly detection. In: IEEE Symposium on Security and Privacy, Oakland, CA, 2001.

Lee, W., S.J. Stolfo, and K.W. Mok. A data mining framework for building intrusion detection models. In: Proceedings of the IEEE Symposium on Security and Privacy, Oakland, CA, 1999, pp. 120–132.

Leung, K. and C. Leckie. Unsupervised anomaly detection in network intrusion detection using clusters. In: Proceedings of the Twenty-Eighth Australasian Conference on Computer Science, Newcastle, Australia, 2005, pp. 333–342.

Li, X., F. Bian, M. Crovella, C. Diot, R. Govindan, G. Iannaccone, and A. Lakhina. Detection and identiﬁcation of network anomalies using sketch subspaces. In: Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement, New York, 2006, pp. 147–152.

Liao, Y.H. and V.R. Vemuri. Use of k-nearest neighbor classiﬁer for intrusion detection. Computers & Security 21 (5) (2002): 439–448.

Liu, Z., G. Florez, S.M. Bridges. A comparison of input representations in neural networks: a case study in intrusion detection. In: Proceedings of the 2002 International Joint Conference on Neural Networks, Honolulu, HI, 2002.

Luo, J. and S.M. Bridges. Mining fuzzy association rules and fuzzy frequency episodes for intrusion detection. International Journal of Intelligent Systems 15 (8) (2000): 687–703.

Mahoney, M.V. and P.K. Chan. Learning nonstationary models of normal network traﬃc for detecting novel attacks. In: Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Canada, 2002, pp. 376–386.

Mannila, H. and H. Toivonen. Discovering generalized episodes using minimal occurrences. In: Proceedings of the 2nd International Conference on Knowledge Discovery in Databases and Data Mining, Portland, OR, 1996.

Platt, J. Fast training support vector machines using sequential minimal optimization. In: Advanced in Kernel Methods-Support Vector Learning, edited by B. Scholkopf, C.J.C. Burges, and A.J. Smola, pp. 185–208. Cambridge, MA: MIT Press, 1999.

Portnoy, L., E. Eskin, and S. Stolfo. Intrusion detection with unlabeled data using clustering. In: Proceedings of ACM CSS Workshop on Data Mining Applied to Security (DMSA), Philadelphia, PA, 2001.

Qiao, Y., X.W. Xin, Y. Bin, and S. Ge. Anomaly intrusion detection method based on HMM. Electronics Letters 38 (13) (2002): 663–664.

Ramaswamy, S., R. Rastogi, and K. Shim. Efficient algorithms for mining outliers from large data sets. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, New York, 2000, pp. 427–438.

Ringberg, H., A. Soule, J. Rexford, and C. Diot. Sensitivity of PCA for traffic anomaly detection. ACM SIGMETRICS Performance Evaluation Review 35 (1) (2007): 109–120.

Sarasamma, S.T. and Q.A. Zhu. Min-max hyperellipsoidal clustering for anomaly detection in network security. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 36 (4) (2006): 887–901.

Soule, A., K. Salamatian, and N. Taft. Combining filtering and statistical methods for anomaly detection. In: Proceedings of the 5th ACM SIGCOMM Conference on Internet Measurement, Berkeley, CA, 2005.

Sugato, B., B. Mikhail, and R.J. Mooney. A probabilistic framework for semi-supervised clustering. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, 2004.

Traore, I. and W. Lu. Unsupervised anomaly detection using an evolutionary extension of k-means algorithm. International Journal of Information and Computer Security 2 (2) (2008): 107–139.

Wang, W., X. Guan, X. Zhang, and L. Yang. Profiling program behavior for anomaly intrusion detection based on the transition and frequency property of computer audit data. Computers & Security 25 (7) (2006): 539–550.

Warrender, C., S. Forrrest, and B. Pearlmutter. Detecting intrusions using system calls: Alternative data models. IEEE Symposium on Security and Privacy, Oakland, CA, 1999, pp. 133–145.

Yamanishi, K. and J.I. Takeuchi. Discovering outlier filtering rules from unlabeled data: Combining a supervised learner with an unsupervised learner. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Canada, 2001, pp. 389–394.

Yamanishi, K., J.I. Takeuchi, G. Williams, and P. Milne. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. In: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, MA, 2000, pp. 320–324.

Ye, N., X.Y. Li, Q. Chen, S.M. Emran, and M.M. Xu. Probabilistic techniques for intrusion detection based on computer audit data. IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans 31 (4) (2001): 266–274.

Ye, N., S.M. Emran, Q. Chen, and S. Vilbert. Multivariate statistical analysis of audit trails for host-based intrusion detection. IEEE Transactions on Computers 51 (2002): 810–820.

Zhang, J. and M. Zulkernine. Anomaly based network intrusion detection with unsupervised outlier detection. In: IEEE International Conference on Communications, Istanbul, Turkey, 2006.

Zhang, J., M. Zulkernine, and A. Haque. Random-forest-based network intrusion detection systems. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 38 (5) (2008): 649–659.

5 Chapter 5: Machine Learning for Hybrid Detection

Agrawal, R., J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In: Proceedings of SIGMOD, Seattle, WA, 1998, pp. 94–105.

Anderson, D., T. Frivold, and A. Valdes. Next-generation intrusion detection expert system (NIDES)—A summary. Technical Report SRI-CSL-95-07, SRI, 1995.

Barbarra, D., J. Couto, S. Jajodia, L. Popyack, and N. Wu. ADAM: Detecting intrusions by data mining. In: Proceedings of the 2001 IEEE, Workshop on Information Assurance and Security, West Point, NY, 2001, pp. 11–16.

Bashah, N., I.B. Shannugam, and A.M. Ahmed. Hybrid intelligent intrusion detection system. World Academy of Science, Engineering and Technology, 11 (2005): 23–26.

Cuppens, F. and A. Miège. Alert correlation in a cooperative intrusion detection framework. In: IEEE Symposium on Research in Security and Privacy, Oakland, CA, 2002.

Dain, O. and R. Cunningham. Building scenarios from a heterogeneous alert stream. In: Proceedings of the 2001 IEEE Workshop on Information Assurance and Security, West Point, NY, 2001a, pp. 231–235.

Dain, O. and R. Cunningham. Fusing a heterogeneous alert stream into scenarios. In: Proceedings of the 2001, ACM Workshop on Data Mining for Security Applications, Philidelphia, PA, 2001b, pp. 1–13.

Endler, D. Intrusion detection: Applying machine learning to Solaris audit data. In: Proceedings of the 1998 Annual Computer Security Applications Conference (ACSAC), Los Alamitos, CA, 1998, pp. 268–279.

Ghosh, A.K. and A. Schwartzbard. A study in using neural networks for anomaly and misuse detection. In: Proceedings of the Eighth USENIX Security Symposium, Washington, DC, 1999, pp. 141–152.

Hu, W.M., W. Hu, and S. Maybank. AdaBoost-based algorithm for network intrusion detection. IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics 38 (2) (2008): 577–583.

Hwang, K., M. Cai, Y. Chen, and M. Qin. Hybrid intrusion detection with weighted signature generation over anomalous internet episodes. IEEE Transactions on Dependable and Secure Computing 4 (1) (2007): 41–55.

Ilgun, K., R.A. Kemmerer, and P.A. Porras. State transition analysis: A rule-based intrusion detection approach. IEEE Transactions on Software Engineering 21 (3) (1995): 181–199.

Lee, W., S.J. Stolfo, and K.W. Mok. A data mining framework for building intrusion detection models. In: Proceedings of the IEEE Symposium on Security and Privacy, Oakland, CA, 1999, pp. 120–132.

Ning, P., D. Xu, C. Healey, and R.S. Amant. Building attack scenarios through integration of complementary alert correlation method. In: Proceedings of the 11th Annual Network and Distributed System Security Symposium, San Diego, CA, 2004.

Porras, P.A. and P.G. Neumann. EMERALD: Event monitoring enabling responses to anomalous live disturbances. In: Proceedings of the Nineteenth National Computer Security, Baltimore, MD, 1997, pp. 353–365.

Qin, M. and K. Hwang. Frequent episode rules for internet tra¢c analysis and anomaly detection. In: Proceedings of IEEE Network Computing and Applications (NAC), Cambridge, MA, 2004.

Seleznyov, A. and S. Puuronen, HIDSUR: A hybrid intrusion detection system based on real-time user recognition. In: Proceedings of 11th International Workshop on Database and Expert Systems Applications, Greenwich, U.K., 2000, pp. 41–45.

Tombini, E., H. Debar, L. Me, and M. Ducasse. A serial combination of anomaly and misuse IDSes applied to HTTP tra¢c. In: Proceedings of Twentieth Annual Computer Security Applications Conference, Tucson, AZ, 2004, pp. 428–437.

Zhang, J. and M. Zulkernine. A hybrid network intrusion detection technique using random forests. In: Proceedings of the First International Conference on Availability, Reliability and Security, Vienna, Austria, 2006a, pp. 262–269.

Zhang, J. and M. Zulkernine. Anomaly based network intrusion detection with unsupervised outlier detection. In: IEEE International Conference on Communications, Istanbul, Turkey, 2006b.

Zhang, J., M. Zulkernine, and A. Haque. Random-forest-based network intrusion detection systems. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 38 (5) (2008): 649–659.

Braynov, S. and M. Jadliwala. Detecting malicious groups of agents. In: Proceedings of the First IEEE Symposium on Multi-Agent Security and Survivability, Philadelphia, PA, 2004.

Cohen, W.W. Fast effective rule induction. In: Proceedings of the 12th International Conference on Machine Learning, San Mateo, CA, 1995.

Conti, G. and K. Abdullah. Passive visual fingerprinting of network attack tools. In: Proceedings of 2004 CCS Workshop on Visualization and Data Mining for Computer Security, Washington, DC, 2004, pp. 45–54.

Gates, C., J.J. McNutt, J.B. Kadane, and M.I. Kellner. Scan detection on very large networks using logistic regression modeling. In: Proceedings of the 11th IEEE Symposium on Computers and Communications (ISCC), Cagliari, Sardin, 2006.

Jung, J., V. Paxson, A.W. Berger, and H. Balakrishnan. Fast portscan detection using sequential hypothesis testing. In: IEEE Symposium on Security and Privacy, Oakland, CA, 2004.

Leckie, C. and Kotagiri, R. A probabilistic approach to detecting network scans. In: Proceedings of the 2002 IEEE Network Operations and Management Symposium, Florence, Italy, 2002, pp. 359–372.

Muelder, C., L. Chen, R. Thomason, K.L. Ma, and T. Bartoletti. Intelligent classification and visualization of network scans. In: Proceedings of the Workshop on Visualization for Computer Security, Sacramento, CA, 2007.

Muelder, C., K.L. Ma, and T. Bartoletti. A visualization methodology for characterization of network scans. In: IEEE Workshops on Visualization for Computer Security, Minneapolis, MN, 2005, pp. 29–38.

Paxson, V. Bro: A system for detecting network intruders in real-time. In Proceedings of the Seventh USENIX Security Symposium, San Antonio, TX, 1998.

Robertson, S., E.V. Siegel, M. Miller, and S.J. Stolfo. Surveillance detection in high bandwidth environments. In:

Proceedings of the 2003 DARPA DISCEX III Conference, Washington, DC, 2003, pp. 130–139.

Roesch, M. Snort-lightweight intrusion detection for networks. In: Proceedings of the 13th USENIX Conference on System Administration, Seattle, WA, 1999, pp. 229–238.

Simon, G., H. Xiong, E. Eilertson, and V. Kumar. Scan detection: A data mining approach. In: Proceedings of the Sixth SIAM International Conference on Data Mining (SDM), Bethesda, MD, 2006, pp. 118–129.

Staniford, S., J.A. Hoagland, and J.M. McAlerney. Practical automated detection of stealthy portscans. In: Proceedings of the Seventh ACM Conference on Computer and Communications Security, Athens, Greece, 2002a.

Staniford, S., J.A. Hoagland, and J.M. McAlerney. Practical automated detection of stealthy portscans. Journal of Computer Security 10, 105–136 (2002b).

Staniford-Chen, S., S. Cheung, R. Crawford, M. Dilger, J. Frank, J. Hoagland, K. Levitt, C. Wee, R. Yip, and D. Zerkle. GrIDS: A graph-based intrusion detection system for large networks. In: The 19th National Information Systems Security Conference, Baltimore, MD, 1996.

Yegneswaran, V., P. Barford, and J. Ullrich. Internet intrusions: Global characteristics and prevalence. In: Proceedings of the 2003 ACM Joint International Conference on Measurement and Modeling of Computer Systems, San Diego, CA, 2003, pp. 138–147.

7 Chapter 7: Machine Learning for Profiling Network Traffic

Apiletti, D., E. Baralis, T. Cerquitelli, and V. D’Elia. Characterizing network tra¢c by means of the NetMine framework. Computer Networks 53 (6) (2008): 774–789.

Chandola, V., E. Banerjee et al. Data mining for cyber security. In: Data Warehousing and Data Mining Techniques for Computer Security, edited by A. Singhal. New York: Springer, 2006.

Cheeseman, P. and J. Strutz. Bayesian classification (autoclass): Theory and results. In: Advances in Knowledge Discovery and Data Mining, edited by G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy U.M. Fayyad. Menlo Park, CA: AAI/MIT Press, 1996.

Erman, J., M. Arlitt, and A. Mahanti. Tra¢c classification using clustering algorithms. In: Proceedings of the 2006 ACM SIGCOMM Workshop on Mining Network Data, Pisa, Italy, 2006.

Ertöz, L., M. Steinbach, and V. Kumar. Finding clusters of di®erent sizes, shapes, and densities in noisy, high dimensional data. In: Proceedings of the Third SIAM International Conference on Data Mining, San Francisco, CA, 2003, pp. 47–58.

Estan, C., S. Savage, and G. Varghese. Automatically inferring patterns of resource consumption in network tra¢c. In: Proceedings of ACMSIGCOMM, Karlsruhe, Germany, 2003, pp. 137–148.

Ester, M., H. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In: The Second International Conference on Knowledge Discovery and Data Mining (KDD), Portland, OR, 1996.

Karagiannis, T., K. Papagiannaki, and M. Faloutsos. BLINC: Multilevel tra¢c classification in the dark. In: Proceedings of ACM SIGCOMM, Philadelphia, PA, 2005, pp. 229–240.

Lakhina, A., M. Crovella, and C. Diot. Characterization of network-wide anomalies in tra¢c ¤ows. In: Proceedings of the Fourth ACM SIGCOMM Conference on Internet Measurement, Taormina, Sicily, Italy, 2004, pp. 201-206.

Lakhina, A., M. Crovella, and C. Diot. Mining anomalies using tra¢c feature distributions. In: Proceedings of the 2005 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, Philadelphia, PA, 2005.

McGregor, A., M. Hall, P. Lorier, and J. Brunskill. Flow clustering using machine learning techniques. In: PAM 2004, Antibes Juan-les-Pins, France, 2004.

Xu, K., X.L. Zhang, and S. Bhattachayya. Internet tra¢c behavior pro�ling for network security monitoring. IEEE/ACM Transactions on Networking (TON) 16 (6) (2008): 1241-1252.

8 Chapter 8: Privacy-Preserving Data Mining

Agrawal, R. and R. Srikant. Privacy-preserving data mining. In: Proceedings of the ACM SIGMOD Conference on Management of Data, Dallas, TX, 2000, pp. 439–450.

Aggarwal, C.C. and P.S. Yu. Privacy-Preserving Data Mining: Models and Algorithms. New York: Springer, 2008.

Barni, M., C. Orlandi, and A. Piva. A privacy-preserving protocol for neural-networkbased computation. In: Proceedings of the Eighth Workshop on Multimedia and Security, Geneva, Switzerland, 2006, pp. 146–151.

Bertina, E., I. Nai Fovino, and L.P. Provenza. A framework for evaluating privacy preserving data mining algorithms. Data Mining and Knowledge Discovery 11 (2) (2005): 121–154.

Dinur, I. and K. Nissim. Revealing information while preserving privacy. In: Proceedings of the 22nd Symposium on Principles of Database Systems (PODS), San Diego, CA, 2003, pp. 202–210.

Du, W. and M.J. Atallah. Secure multi-party computation problems and their applications: A review and open problems. In: Proceedings of New Security Paradigms Workshop, Cloudcroft, NM, 2001, pp. 11–20.

Du, W. and Z. Zhan. Building decision tree classifier on private data. In: Proceedings of the IEEE ICDM Workshop on Privacy, Security and Data Mining, Maebashi City, Japan, 2002.

Du, W., Y.S. Han, and S. Chen. Privacy-preserving multivariate statistical analysis: Linear regression and classification. In: Proceedings of SIAM International Conference on Data Mining (SDM), Nashville, TN, 2004.

Dwork, C. and S. Yekhanin. New efficient attacks on statistical disclosure control mechanisms. In: Proceedings of the 28th Annual Conference on Cryptology: Advances in Cryptology, Santa Barbara, CA, 2008, pp. 469–480.

Evfimievski, A., J. Gehrke, and R.J. Srikant. Limiting privacy breaches in privacy preserving data mining. In: Proceedings of the ACM SIGACT SIGMOD SIGART Symposium on Principles of Database Systems, San Diego, CA, 2003, pp. 211–222.

Evfimievski, A., R. Srikant, R. Agrawal, and J. Gehrke.
Privacy preserving mining of association rules. In:
Proceedings of the Eighth ACM SIGKDDD International
Conference on Knowledge Discovery and Data Mining,
Edmonton, Alberta, Canada, 2002.

Evfimievski, A., R. Srikant, R. Agrawal, and J. Gehrke.
Privacy preserving mining of association rules. Information
Systems 29 (4) (2004): 343–364.

Hinke, T.H., H.S. Delugach, and R.P. Wolf. Protecting
databases from inference attacks. Computers and Security
16 (8) (1997): 687–708.

Kantarcioglu, M. and C. Clifton. Privately computing a
distributed k-nn classifier. In: Proceedings of the Eighth
European Conference on Principles and Practice of Knowledge
Discovery in Databases, Pisa, Italy, 2004, pp. 279–290.

Kantarcıoglu, M., J. Jin, and C. Clifton. When do data
mining results violate privacy? In: Proceedings of the
2004 ACM SIGKDD International Conference on Knowledge
Discovery and Data Mining, Seattle, WA, 2008, pp. 599–604.

Online Security and Privacy Study. A report from Harris
Interactive 2009. http://www.

Shannon, C.E. Communication theory of secrecy systems. Bell
System Technical Journal 28 (4) (1949): 656–715.

Sweeney, L. K-anonymity: A model for protecting privacy.
International Journal of Uncertainty, Fuzziness and
Knowledge-Based Systems 10 (5) (2002): 557–570.

Vaidya, J. and C. Clifton. Privacy-preserving K-means
clustering over vertically partitioned data. In:
Proceedings of the Ninth ACM SIGKDD International
Conference on Knowledge Discovery and Data Mining,
Washington, DC, 2003.

Vaidya, J. and C. Clifton. Privacy-preserving data mining:
Why, how, and when. IEEE Security and Privacy 2 (2004):
19–27.

Verykios, V.S., E. Bertino, I.N Fovino, L.P. Provenza, Y,
Saygin, and Y. Theodoridis. Stateof-the-art in privacy
preserving data mining. ACM SIGMOD Record 33 (1) (2004a):
50–57.

Verykios, V.S., A. Elmagamid, E. Bertino, Y. Saygin, and E.

Dasseni. Association rule hiding. IEEE Transactions on Knowledge and Data Engineering 16 (4) (2004b): 434–447.

Willenborg, L. and T. DeWaal. Elements of Statistical Disclosure Control, Lecture Notes in Statistics, Vol. 155. New York: Springer, 2001.

Wright, R. and Z. Yang. Privacy-preserving Bayesian network structure computation on distributed heterogeneous data. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, 2004.

Yang, Z. and R.N. Wright. Privacy-preserving computation of Bayesian networks on vertically partitioning data. IEEE Transactions on Knowledge and Data Engineering 18 (9) (2006): 1253–1264.

Yao, A.C. Protocols for secure computations. In: Proceedings of the Third Annual IEEE Symposium on Foundations of Computer Science, Chicago, IL, 1982.

Yu, H., X. Jiang, and J. Vaidya. Privacy-preserving SVM using nonlinear kernels on horizontally partitioned data. In: Proceedings of the 2006 ACM Symposium on Applied Computing, Dijon, France, 2006.

9 Chapter 9: Emerging Challenges in Cybersecurity

Axelsson, S. The base-rate fallacy and its implications for the di¢culty of intrusion detection. ACM Transactions on Information and System Security 3 (2000): 186–205.

Bianchi, G. et al. Towards privacy-preserving network monitoring: Issues and challenges. In: The 18th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, Athens, Greece, 2007.

Bianchi, G., S. Teofili, and M. Pomposini. New directions in privacy-preserving anomaly detection for network tra¢c. In: Proceedings of the First ACM Workshop on Network Data Anonymization, Alexandria, VA, 2008, pp. 11–18.

Data Loss Prevention Best Practices: Managing Sensitive Data in the Enterprise. A report from IronPort Systems, San Bruno, CA, 2007.

He, H. and E.A. Garcia. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering 21 (9) (2009): 1263–1284.

McHugh, J. Testing intrusion detection systems: A critique of the 1998 and 1999 DARPA intrusion detection. ACM Transactions on Information and System Security 3 (2000): 262–294.

Messmer, E. America's 10 most wanted botnets. Damballa, Atlanta, GA, 2009.

Mustaque, A., A. Dave et al. Emerging cyber threats report for 2009. Georgia Tech Information Security Center, 2008 GTISC security summit—Emerging cybersecurity threats.

Pokrajec, D., A. LAzarevic, and L.J. Latecki. Incremental local outlier detection for data streams. In: Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining, Honolulu, HI, 2007.

PRIvacy-Aware Secure Monitoring. http://fp7-prism.eu/index.php?option=com_content&task=view&id=20&Itemid=29 (accessed 2010).

Stolfo, S.J., W. Fan, and W. Lee. Cost-based modeling for fraud and intrusion detectors: results from the JAM project. In: DARPA Information Survivability Conference & Exposition, Hilton Head, SC, 2000, pp. 120–144.

Tikk, E., K. Kaska, K. Rünnimeri, M. Kert, A.M. Talihärm, and L. Vihul. Cyber Attacks against Georgia: Legal Lessons Identified. NATO, 2008.

Valdya, J. and C. Clifton. Privacy-preserving outlier detection. In: Proceedings of the Fourth IEEE International Conference on Data Mining, Brighton, U.K., 2004, pp. 233–240.

Virtual Criminology Report 2009: Virtually Here: The Age of Cyber Warfare. McAfee, Santa Clara, CA, 2009.

Zhu, Z., G. Lu, Y. Chen, Z. Fu, P. Roberts, and K. Han. Botnet research survey. In: Annual IEEE International Computer Software and Application Conference, Turku, Finland, 2008, pp. 967–973.

Zimmermann, J. and G. Mohay. Distributed intrusion detection in clusters based on noninterference. In: Proceedings of the Australasian Workshops on Grid Computing and E-Research, Hobart, Tasmania, Australia, 2006, pp. 89–95.