APPLICATIONS OF DATA MINING IN COMPUTER SECURITY

ADVANCES IN INFORMATION SECURITY

Sushil Jajodia
Consulting editor

Center for Secure Information Systems
George Mason University
Fairfax, VA 22030-4444

email: jajodia@gmu.edu

Additional titles in the series:

MOBILE COMPUTATION WITH FUNCTIONS by Zeliha Dilsun Kırlı, ISBN: 1-4020-7024-1

TRUSTED RECOVERY AND DEFENSIVE INFORMATION WARFARE by Peng Liu and Sushil Jajodia, ISBN: 0-7923-7572-6

RECENT ADVANCES IN RSA CRYPTOGRAPHY *by Stefan Katzenbeisser, ISBN: 0-7923-7438-X*

E-COMMERCE SECURITY AND PRIVACY *by Anup K. Ghosh, ISBN: 0-7923-7399-5*

INFORMATION HIDING: Steganography and Watermarking-Attacks and Countermeasures by Neil F. Johnson, Zoran Duric, and Sushil Jajodia, ISBN: 0-7923-7204-2

Additional information about this series can be obtained from
www.wkap.nl/series.htm/ADIS.

APPLICATIONS OF DATA MINING IN COMPUTER SECURITY

edited by

Daniel Barbará
Sushil Jajodia
George Mason University
U.S.A.

SPRINGER SCIENCE+BUSINESS MEDIA, LLC

ISBN 978-1-4613-5321-8 ISBN 978-1-4615-0953-0 (eBook)
DOI 10.1007/978-1-4615-0953-0

Library of Congress Cataloging-in-Publication Data

A C.I.P. Catalogue record for this book is available
from the Library of Congress.

Copyright © 2002 by Springer Science+Business Media New York
Originally published by Kluwer Academic Publishers in 2002
Softcover reprint of the hardcover 1st edition 2002

All rights reserved. No part of this work may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, microfilming, recording, or otherwise, without the written permission from the Publisher, with the exception of any material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work.

Permission for books published in Europe: permissions@wkap.nl
Permissions for books published in the United States of America: permissions@wkap.com

Printed on acid-free paper.

Series Foreword

ADVANCES IN INFORMATION SECURITY

Sushil Jajodia
Consulting Editor

Center for Secure Information Systems
George Mason University
Fairfax, VA 22030-4444

email: jajodia@gmu.edu

Welcome to the sixth volume of the Kluwer International Series on ADVANCES IN INFORMATION SECURITY. The goals of this series are, one, to establish the state of the art of, and set the course for future research in information security and, two, to serve as a central reference source for advanced and timely topics in information security research and development. The scope of this series includes all aspects of computer and network security and related areas such as fault tolerance and software assurance.

ADVANCES IN INFORMATION SECURITY aims to publish thorough and cohesive overviews of specific topics in information security, as well as works that are larger in scope or contain more detailed background information than can be accommodated in shorter survey articles. The series also serves as a forum for topics that may not have reached a level of maturity to warrant a comprehensive textbook treatment.

The success of this series depends on contributions by researchers and developers such as you. If you have an idea for a book that is appropriate for this series, I encourage you to contact me. I would be happy to discuss any potential projects with you. Additional information about this series can be obtained from www.wkap.nl/series.htm/ADIS.

About This Volume

This sixth volume of the series is entitled*APPLICATIONS OF DATA MINING IN COMPUTER SECURITY*, edited by Daniel Barbarà and Sushil Jajodia.

Computer intrusions are becoming commonplace and outpacing our capacity to detect, analyze, and counteract them. Since intrusions usually leave traces in the audit data trails, it is only natural to think about this problem in a data-centered way. Some research groups have been successfully using data mining techniques for effectively implementing tools to detect and analyze intrusions.

This volume offers nine articles from leading researchers; eight of these articles focus on the use of data mining for intrusion detection, including one that surveys the state of modern intrusion detection using data mining approaches and another that critically examines these approaches. The last article deals with the application of data mining to computer forensics. Collectively, these articles provide a comprehensive summary of current findings in this fruitful research field.

SUSHIL JAJODIA
Consulting Editor

Contents

List of Figures xiii

List of Tables xvii

Preface xix

1

Modern Intrusion Detection, Data Mining, and Degrees of Attack Guilt 1

Steven Noel, *Duminda Wijesekera* and *Charles Youman*

1. Introduction 2
2. Detection Strategies 3
   - 2.1 Misuse Detection 4
   - 2.1.1 Expert Systems 4
   - 2.1.2 Signature Analysis 5
   - 2.1.3 State-Transition Analysis 6
   - 2.1.4 Data Mining 7
   - 2.1.5 Other Approaches 8
   - 2.2 Anomaly Detection 8
   - 2.2.1 Statistical Methods 9
   - 2.2.2 Expert Systems 10
   - 2.2.3 Data Mining 10
   - 2.2.4 Other Approaches 12
3. Data Sources 12
4. Degrees of Attack Guilt 14
   - 4.1 Misuse Detection 15
   - 4.1.1 Knowledge-Based Methods 16
   - 4.1.2 Machine-Learning Methods 17
   - 4.2 Anomaly Detection 18
   - 4.2.1 Knowledge-Based Methods 18
   - 4.2.2 Statistical Methods 19
   - 4.2.3 Machine-Learning Methods 20
5. Conclusion 25

References 25

2

Data Mining for Intrusion Detection 33

Klaus Julisch
1. Introduction 33
2. Data Mining Basics 34
2.1 Data Mining, KDD, and Related Fields 34
2.2 Some Data Mining Techniques 36
2.2.1 Association Rules 37
2.2.2 Frequent Episode Rules 38
2.2.3 Classification 39
2.2.4 Clustering 40
2.3 Research Challenges in Data Mining 40
3. Data Mining Meets Intrusion Detection 41
3.1 MADAM ID 43
3.2 ADAM 45
3.3 Clustering of Unlabeled ID Data 46
3.4 Mining the Alarm Stream 47
3.5 Further Reading 49
4. Observations on the State of the Art 50
4.1 Data Mining, but no Knowledge Discovery 50
4.2 Disregard of Other KDD Steps 51
4.3 Too Strong Assumptions 52
4.4 Narrow Scope of Research Activities 53
5. Future Research Directions 54
6. Summary 56

References 57

3
An Architecture for Anomaly Detection 63
Daniel Barbará, *Julia Couto*, *Sushil Jajodia* and *Ningning Wu*
1. Introduction 63
2. Architecture 65
2.1 Filter 65
2.2 Profile 67
2.3 Profile Builder 67
2.4 Diagnoser 67
3. ADAM: an implementation of the architecture 67
4. Experiences 72
5. Breaking the dependency on training data 73
6. Future 74

References 75

4
A Geometric Framework for Unsupervised Anomaly Detection 77
Eleazar Eskin, *Andrew Arnold*, *Michael Prerau*, *Leonid Portnoy* and *Sal Stolfo*
1. Introduction 78
2. Unsupervised Anomaly Detection 81
3. A Geometric Framework for Unsupervised Anomaly Detection 83
3.1 Feature Spaces 83
3.2 Kernel Functions 84
3.3 Convolution Kernels 85

4. Detecting Outliers in Feature Spaces 85
5. Algorithm 1: Cluster-based Estimation 86
6. Algorithm 2: K-nearest neighbor 87
7. Algorithm 3: One Class SVM 89
8. Feature Spaces for Intrusion Detection 91
   8.1 Data-dependent Normalization Kernels 92
   8.2 Kernels for Sequences: The Spectrum Kernel 92
9. Experiments 93
   9.1 Performance measures 93
   9.2 Data Set Descriptions 94
   9.3 Experimental Setup 95
   9.4 Experimental Results 96
10. Discussion 98

References 99

5
Fusing a Heterogeneous Alert Stream into Scenarios 103
Oliver Dain and *Robert K. Cunningham*
1. Introduction 104
2. Fusion Approach 105
3. Architecture 106
4. Definitions 107
5. Probability Assignment 108
   5.1 Data Sources and Use 108
   5.2 Naïve Technique 111
   5.3 Heuristic Technique 112
   5.4 Data Mining Techniques 114
6. Experimental Results 115
   6.1 Naïve Technique 116
   6.2 Heuristic Technique 117
   6.3 Data Mining Techniques 117
7. System Benefits 119
8. Discussion and Summary 120

References 120

6
Using MIB II Variables for Network Intrusion Detection 123
Xinzhou Qin, *Wenke Lee*, *Lundy Lewis* and *João B. D. Cabrera*
1. Introduction 124
2. Background 125
   2.1 MIB II 125
   2.2 Entropy and Conditional Entropy 126
3. Model Construction 127
   3.1 Model Architecture 127
   3.2 Anomaly Detection Module 129
   3.2.1 Anomaly Detection Model Design Overview 129
   3.2.2 Anomaly Detection Module Construction 129
4. Experiments and Performance Evaluation 134

4.1 Normal Data Sets 134
4.2 Evaluation under Attacks 135
4.2.1 Misuse Detection 135
4.2.2 Anomaly Detection 140
5. Discussion 146
6. Related Work 148
7. Conclusions and Future Work 149

References 149

7
Adaptive Model Generation 153
Andrew Honig, *Andrew Howard*, *Eleazar Eskin* and *Sal Stolfo*
1. Introduction 154
2. Components of Adaptive Model Generation 157
2.1 Real Time Components 159
2.2 Data Warehouse 163
2.3 Detection Model Management 165
2.4 Data Analysis Engines 167
2.5 Efficiency consideration 174
3. Capabilities of Adaptive Model Generation 175
3.1 Real Time Detection Capabilities 175
3.2 Automatic Data Collection and Data Warehousing 175
3.3 Model Generation and Management 176
3.4 Data Analysis Capabilities 176
3.5 Correlation of Multiple Sensors 178
4. Model Generation Algorithms 179
4.1 Misuse Detection 179
4.2 Anomaly Detection 179
4.3 Unsupervised Anomaly Detection 180
5. Model Generation Example: SVM 180
5.1 SVM Algorithm 181
5.2 SVM for Misuse Detection in AMG 182
5.3 Unsupervised SVM Algorithm 183
5.4 Unsupervised SVM for Unsupervised Anomaly Detection 184
6. System Example 1: Registry Anomaly Detection 185
6.1 The RAD Data Model 185
6.2 The RAD Sensor 185
6.3 The RAD Classification Algorithm 186
6.4 The RAD Detector 187
7. System Example 2: HAUNT 187
7.1 HAUNT Sensor 188
7.2 HAUNT Classification Algorithm 188
7.3 HAUNT Detector 188
7.4 HAUNT Feature Extraction 189
8. Conclusion 190

References 191

8
Proactive Intrusion Detection 195

João B. D. Cabrera, *Lundy Lewis*, *Xinzhou Qin*, *Wenke Lee* and *Raman K. Mehra*

1. Introduction 196
2. Information Assurance, Data Mining, and Proactive Intrusion Detection 198
   2.1 Intrusion Detection Systems 198
   2.2 A Thought Experiment 198
   2.3 Proactive Intrusion Detection 204
3. A methodology for discovering precursors - Assumptions, Objectives, Procedure and Analysis 206
   3.1 Notation and Definitions 206
   3.1.1 Time Series, Multivariate Time Series and Collections 206
   3.1.2 Events, Event Sequences, Causal Rules and Precursor Rules 207
   3.2 Assumptions, Problem Set-Up, Objectives and Procedure 208
   3.3 Analysis - Detection and Gradation of Causality in Time Series 211
   3.3.1 Notation and Definitions 211
   3.3.2 The Granger Causality Test as an Exploratory Tool 212
   3.3.3 GCT and the Extraction of Precursor Rules - Modeling and Theoretical Developments 213
4. A Case Study - Precursor Rules for Distributed Denial of Service Attacks 217
   4.1 DDoS Attacks and the experiments 217
   4.2 TFN2K Ping Flood - Extracting Precursor Rules 219
5. Conclusions 222

References 223

9

E-mail Authorship Attribution for Computer Forensics 229

Olivier de Vel, *Alison Anderson*, *Mal Corney* and *George Mohay*

1. Introduction and Motivation 230
   1.1 Computer Forensics 230
   1.2 E-mail Forensics 232
2. Authorship Attribution 234
3. E-mail Authorship Attribution 238
4. Support Vector Machine Classifier 239
5. E-mail Corpus and Methodology 240
6. Results and Discussion 244
7. Conclusions 246

References 247

List of Figures

1.1 General Degrees of Attack Guilt 15

1.2 Degrees of Attack Guilt for Knowledge-based Misuse Detection 16

1.3 Degrees of Attack Guilt for Machine-learning Misuse Detection 17

1.4 Degrees of Attack Guilt for Knowledge-based Anomaly Detection 18

1.5 Degrees of Attack Guilt for Statistical Anomaly Detection 19

1.6 Combining data mining and classification for anomaly detection 21

1.7 Degrees of Guilt for the Training Phase 22

1.8 Degrees of Guilt for Connection Mining during Detection 23

1.9 Degrees of Guilt for Attack Classification during Detection 23

1.10 Degrees of Guilt with respect to Attack Classification Confidence 24

1.11 Degrees of Guilt for Overall Machine-learning Approach 24

2.1 Data mining process of building misuse detection systems. 44

3.4 Results of using ADAM in the 1999 Lincoln Labs competition data. 73

5.1 (a) Possible scenario assignments for a new alert, C, given 2 existing alerts, A, and B. (b) Possible scenario assignments for a new alert, D, given 3 existing alerts, A, B, and C. 106

5.2 Alert occurrence times in the DEF CON data. Each horizontal line depicts a single scenario. The points on the line indicate when an alert occurred. The symbol indicates the class of the alert. Depicted are the twenty five scenarios containing the largest number of alerts. 109

5.3 Decision surface for a single transition type. The height of the curve is the probability that two alerts in the given time span with the given r value belong in the same scenario. The black plane is the threshold below which the new alert would not join the scenario and would start a new scenario. 113

6.1 MIB II-based ID Architecture 128

6.2 MIB II-based Anomaly ID Model 130

6.3 Conditional Entropy of MIB II object *ipInReceives* 132

6.4 Misclassification rate of MIB II object *ipInReceives* 132

6.5 Accuracy over Cost of MIB II object *ipInReceives* 133

6.6 *icmpInEchos* under normal condition 136

6.7 *icmpInEchos* under Ping Flood attack 136

6.8 *udpInDatagrams* under normal condition 138

6.9 *udpInDatagrams* under UDP Flood attack 138

6.10 *udpInErrors* under normal condition 139

6.11 *udpInErrors* under UDP Flood attack 139

6.12 *ICMP_In* ID sub-module under normal condition 142

6.13 *ICMP_In* ID sub-module under Mix Flood attack 142

6.14 *ICMP_Out* ID sub-module under normal condition 143

6.15 *ICMP_Out* ID sub-module under Mix Flood attack 143

6.16 *UDP_In_Error* ID sub-module under normal condition 144

6.17 *UDP_In_Error* ID sub-module under Mix Flood attack 144

6.18 *TCP_In* ID sub-module under normal condition 145

6.19 *TCP_In* ID sub-module under Mix Flood attack 145

7.1 The AMG System Architecture 158

7.2 Visualization of Data in Database 168

7.3 Visualization of SQL Query 169

8.1 Authentication, Passive Intrusion Detection and Proactive Intrusion Detection. Alarm $A0$ is connected to an Authenticator. Alarm $A1$ enables Passive Intrusion Detection, while alarms $A2$ and $A3$ enable Proactive Intrusion Detection. 199

8.2 Timeline of the outputs of the alarms. It is assumed that the time elapsed between the entry of a malicious agent in the house and malicious activity is negligible. 202

8.3 Proactive Intrusion Detection - Extracting Temporal Rules. 205

8.4 The idealized inputs and output signals. p is the length of the window used for parameter estimation when applying GCT. If $p \geq r$, the representation (8.1) captures model (8.4) exactly, and the **Precursor ⇒ Phenomenon** rule is "visible" through the model. When $H(q^{-1}) = 1$, the response in y collapses into a blip. 215

8.5 DDoS Attacks - A simplified Timeline. 218

8.6 TFN2K Ping Flood: Selected MIB variables at the Attacker and at the Target. 219

List of Tables

2.1 Sample database table. 37

4.1 Lincoln Labs Data Summary 95

4.2 Selected points from the ROC curves of the performance of each algorithm over the KDD Cup 1999 Data. 98

5.1 r value calculation for addresses 172.16.112.20 and 172.16.112.40 107

5.2 Confusion matrices from Naïve approach on test data set . Two hundred forty six of the 4,041 patterns that should have produced a "join" decision (6.09%) incorrectly produced a "don't join" decision, and ten of the 246,320 patterns that should have produced a "don't join" decision incorrectly produced a "join" decision. 116

5.3 Confusion matrices from heuristic approach on test data set. The algorithm tries to match the human decision. For example, 4,041 test examples should have produced a "join" decision. The algorithm correctly produced the "join" decision 3,589 times (88.81% of the time). 117

5.4 Confusion matrices from a decision tree on the test data set. The decision tree tries to match the human decision. For example, 4,041 examples should have produced a "join" decision. The decision tree produced the correct decision on 4,037 of these (99.90% of the time). 118

6.1 Misuse Detections by MIB II-based ID Model 140

6.2 Anomaly Detections by MIB II-based ID Model 147

8.1 Key Variables at the Attacker for TFN2K - Ground Truth. 221

8.2 TFN2K Ping Flood Run 1: Top MIBs at the Attacker according to the g statistic. 221

8.3 Results of Step 2: Detection Rates and FA Rates for MIB variables that contain precursors to DDoS Attacks. 222

8.4 Final Results: Detection Rates and FA Rates for Events at MIB variables for TFN2K Ping Flood. 222

9.1 Summary statistics of the e-mail newsgroup and author corpus used in the experiment. 241

9.2 E-mail document body style marker attributes. Total of 170 features are used in the experiment. See text for clarification. 242

9.3 E-mail document body structural attributes. Total of 21 attributes/features are used in the experiment. See text for clarification. 243

9.4 Per-author-category P_{AC_i}, R_{AC_i} and F_{1,AC_i} categorisation performance results (in %) for the four different author categories ($i = 1, \ldots, 4$). The newsgroup `aus.tv` is used as the training set (see text).^{a} 245

Preface

Data mining is becoming a pervasive technology in activities as diverse as using historical data to predict the success of a marketing campaign, looking for patterns in financial transactions to discover illegal activities, or analyzing genome sequences. From this perspective, it was just a matter of time for the discipline to reach the important area of computer security. This book presents a collection of research efforts on the use of data mining in computer security.

Data mining has been loosely defined as the process of extracting information from large amounts of data. In the context of security, the information we are seeking is the knowledge of whether a security breach has been experienced, and, if the answer is yes, who is the perpetrator. This information could be collected in the context of discovering intrusions that aim to breach the privacy of services, or data in a computer system or, alternatively, in the context of discovering evidence left in a computer system as part of a criminal activity.

This collection concentrates heavily on the use of data mining in the area of intrusion detection. The reason for this is twofold. First, the volume of data dealing with both network and host activity is so large that it makes it an ideal candidate for using data mining techniques. Second, intrusion detection is an extremely critical activity. To understand this it is enough to look at the current statistics. Ten major government agencies accounting for 99% of the federal budget have been compromised in the recent past. In the year 2000, a massive, coordinated attack successfully brought down some of the major e-commerce web sites in the United States. Moreover, it is estimated that less than 4% of the attacks are actually detected or reported. As a society, we have become extremely dependent of the use of information systems, so much so that the danger of serious disruption of crucial operations is frightening. As a result, it is no surprise that researchers have produced a relatively large volume of work in the area of data mining in support of intrusion detection.

The rest of the work presented in this volume addresses the application of data mining to an equally pressing area: computer forensics. This area has widened recently to address activities such as law enforcement using digital evidence. Although the amount of work is not as large as in intrusion detection, computer forensics proves to be a fruitful arena for research in data mining techniques.

Data mining holds the promise of being an effective tool to help security activities and, in some sense, the proof of its applicability can be found in the pages of this book. However, there is still a long road to travel and we hope that this volume will inspire researchers and practitioners to undertake some steps in this direction.

Acknowledgments

We are extremely grateful to authors for their contributions and to Jia-Ling Lin who assisted with every aspect of this book, ranging from collecting of manuscripts to dealing with all matters related to Kluwer style files. It is also a pleasure to acknowledge Joe Giordano, Brian Spink, and Leonard Popyack of the Air Force Research Laboratory/Rome for their support of our research in the application of data mining to intrusion detection.

DANIEL BARBARÁ

SUSHIL JAJODIA

FAIRFAX, VA