

Contents

I	Prologue	1
1	Introduction	3
1.1	Problem Description	4
1.2	Contribution	6
1.3	Outline of the Thesis	8
2	Background	9
2.1	Flow Capturing and Intrusion Detection	10
2.1.1	The Flow Standard IPFIX	10
2.1.2	A Taxonomy of Intrusion Detection Systems	13
2.2	Machine Learning	14
2.2.1	Supervised and Unsupervised Learning	14
2.2.2	Model Transparency	19
2.2.3	Nonstationary Environments	21
2.3	Big Data Processing	24
2.3.1	The Phenomenon	24
2.3.2	Reference Architectures	25
2.3.3	Stream Processing Using Flink	27
2.3.4	Parallel Relational Databases	28
II	Uncertainty Management Aspects	31
3	Rudiments of Rough Set Theory	33
3.1	Introduction	34
3.2	Information and Decision Systems	35
3.3	Indiscernibility Relation	36
3.4	Concept Approximation	37
3.5	Reducts and Core	40
3.6	Variable Precision	42
3.6.1	Majority Inclusion	42
3.6.2	Beta Approximation	43
3.7	Summary	45
4	In-Database Variable Precision Rough Sets	47
4.1	Introduction	48
4.2	Combining Rough Sets and Databases	49
4.3	Basic Considerations and Database Notations	50

4.4	Computing Variable Precision Rough Sets	52
4.4.1	Restructuring the Concept Approximation	53
4.4.2	Mapping Variable Precision Rough Sets	54
4.4.3	Deducing Traditional Rough Sets	54
4.4.4	Complexity and Implementation Details	56
4.5	Computing Core and Reducts	58
4.5.1	The Virtue of Imperfection	59
4.5.2	Realization and Complexity	59
4.6	Comparative Study	61
4.7	Discussion and Summary	65
III	Hybrid Flow-based Intrusion Detection	69
5	System Rationale and Design Considerations	71
5.1	Introduction	72
5.2	Threat Model	73
5.3	Operational Aspects	77
5.4	Related Hybridizations	82
5.5	Conceptual Architecture	84
5.6	Summary	87
6	The NDSec-1 Network Traces	89
6.1	Introduction	90
6.2	Common Benchmark Data Sets	91
6.3	Infrastructure and Attack Scenarios	92
6.3.1	Network Infrastructure	93
6.3.2	Attack Scenarios	93
6.3.3	Recap of the Captures	96
6.4	Qualitative Evaluation	96
6.4.1	Comparative Study	97
6.4.2	Practical Insights Using Snort	98
6.5	Discussion and Summary	100
7	Flow Feature Analysis	101
7.1	Introduction	102
7.2	Flow-based Analysis Using NetFlow/IPFIX	103
7.3	Flow Feature Selection	104
7.3.1	Employed Traffic Traces	104
7.3.2	Feature Extraction and Preprocessing	106
7.3.3	Algorithm and Results	109
7.4	Empirical Assessment	113
7.4.1	Learner Categories and Measures	114
7.4.2	Parameter Selection Using ROC Analysis	116
7.4.3	Runtime Incorporation and Consolidation	117
7.4.4	Discussion	121
7.5	Summary	122

8	Anomaly Detection	125
8.1	Introduction	126
8.2	Background	128
8.2.1	Ensemble Learning Preliminaries	128
8.2.2	Adaptive Ensembles	132
8.2.3	Initially Labeled Environments	134
8.3	Detection Component	135
8.3.1	Ensemble Composition	135
8.3.2	Unsupervised Supplements	139
8.3.3	Discussion	143
8.4	Summary	144
9	Pattern Building	147
9.1	Introduction	148
9.2	Related Work	150
9.3	Incremental In-Database Rule Inducing	151
9.3.1	Knowledge Representation	151
9.3.2	Partial Memory	154
9.3.3	Induction Process	156
9.3.4	Implementation Aspects	161
9.4	General Evaluation Under Drifting Conditions	166
9.4.1	Experimental Setup	166
9.4.2	Predictive Capabilities	169
9.4.3	Discovery-oriented Capabilities	177
9.4.4	Time Consumption	181
9.5	Discussion and Summary	183
10	System Deployment and Evaluation	189
10.1	Introduction	190
10.2	System Deployment	190
10.2.1	Main Inspection Pipeline	191
10.2.2	Incorporating the Pattern Building Process	192
10.2.3	Data Preprocessing Steps and Correlations	195
10.2.4	Pattern Matching and Distribution	198
10.2.5	Anomaly Detection and Warm-up	202
10.2.6	Alert Management and Persistence	205
10.2.7	Concluding Picture	205
10.3	System Evaluation	208
10.3.1	Experimental Setup	208
10.3.2	Identifying Operating Points	214
10.3.3	Assessment on Synthetic Drifts	218
10.3.4	Assessment on Enterprise Traces	222
10.3.5	Scalability Potentials	229
10.3.6	Discussion	237
10.4	Summary	242

IV	Epilogue	245
11	Conclusion and Outlook	247
11.1	Conclusion	248
11.2	Outlook	250
	Appendixes	253
	References	259