

Contents

Preface	xiii
Acknowledgments	xv
1	
An Introduction to Outlier Analysis	1
1. Introduction	1
2. The Data Model is Everything	6
3. The Basic Outlier Models	10
3.1 Extreme Value Analysis	10
3.2 Probabilistic and Statistical Models	12
3.3 Linear Models	13
3.4 Proximity-based Models	14
3.5 Information Theoretic Models	16
3.6 High-Dimensional Outlier Detection	18
4. Meta-Algorithms for Outlier Analysis	19
4.1 Sequential Ensembles	20
4.2 Independent Ensembles	21
5. The Basic Data Types for Analysis	22
5.1 Categorical, Text and Mixed Attributes	23
5.2 When the Data Values have Dependencies	23
6. Supervised Outlier Detection	28
7. Outlier Evaluation Techniques	31
8. Conclusions and Summary	35
9. Bibliographic Survey	35
10. Exercises	38
2	
Probabilistic and Statistical Models for Outlier Detection	41
1. Introduction	41
2. Statistical Methods for Extreme Value Analysis	43
2.1 Probabilistic Tail Inequalities	43
2.2 Statistical Tail Confidence Tests	50
3. Extreme Value Analysis in Multivariate Data	54
3.1 Depth-based Methods	55
3.2 Deviation-based Methods	56
3.3 Angle-based Outlier Detection	57
3.4 Distance Distribution-based Methods	60
4. Probabilistic Mixture Modeling for Outlier Analysis	62
5. Limitations of Probabilistic Modeling	68

6.	Conclusions and Summary	69
7.	Bibliographic Survey	70
8.	Exercises	72
3		
	Linear Models for Outlier Detection	75
1.	Introduction	75
2.	Linear Regression Models	78
2.1	Modeling with Dependent Variables	80
2.2	Regression Modeling for Mean Square Projection Error	84
3.	Principal Component Analysis	85
3.1	Normalization Issues	90
3.2	Applications to Noise Correction	91
3.3	How Many Eigenvectors?	92
4.	Limitations of Regression Analysis	94
5.	Conclusions and Summary	95
6.	Bibliographic Survey	95
7.	Exercises	97
4		
	Proximity-based Outlier Detection	101
1.	Introduction	101
2.	Clusters and Outliers: The Complementary Relationship	103
3.	Distance-based Outlier Analysis	108
3.1	Cell-based Methods	109
3.2	Index-based Methods	112
3.3	Reverse Nearest Neighbor Approach	115
3.4	Intensional Knowledge of Distance-based Outliers	116
3.5	Discussion of Distance-based Methods	117
4.	Density-based Outliers	118
4.1	LOF: Local Outlier Factor	119
4.2	LOCI: Local Correlation Integral	120
4.3	Histogram-based Techniques	123
4.4	Kernel Density Estimation	124
5.	Limitations of Proximity-based Detection	125
6.	Conclusions and Summary	126
7.	Bibliographic Survey	126
8.	Exercises	132
5		
	High-Dimensional Outlier Detection: The Subspace Method	135
1.	Introduction	135
2.	Projected Outliers with Grids	140
2.1	Defining Abnormal Lower Dimensional Projections	140
2.2	Evolutionary Algorithms for Outlier Detection	141
3.	Distance-based Subspace Outlier Detection	144
3.1	Subspace Outlier Degree	145
3.2	Finding Distance-based Outlying Subspaces	146
4.	Combining Outliers from Multiple Subspaces	147
4.1	Random Subspace Sampling	147
4.2	Selecting High Contrast Subspaces	149

4.3	Local Selection of Subspace Projections	150
5.	Generalized Subspaces	153
6.	Discussion of Subspace Analysis	159
7.	Conclusions and Summary	162
8.	Bibliographic Survey	163
9.	Exercises	166
6		
	Supervised Outlier Detection	169
1.	Introduction	169
2.	The Fully Supervised Scenario: Rare Class Detection	173
2.1	Cost Sensitive Learning	174
2.2	Adaptive Re-sampling	180
2.3	Boosting Methods	182
3.	The Semi-Supervised Scenario: Positive and Unlabeled Data	184
3.1	Difficult Cases and One-Class Learning	185
4.	The Semi-Supervised Scenario: Novel Class Detection	186
4.1	One Class Novelty Detection	187
4.2	Combining Novel Class Detection with Rare Class Detection	189
4.3	Online Novelty Detection	189
5.	Human Supervision	190
5.1	Active Learning	191
5.2	Outlier by Example	193
6.	Conclusions and Summary	194
7.	Bibliographic Survey	194
8.	Exercises	197
7		
	Outlier Detection in Categorical, Text and Mixed Attribute Data	199
1.	Introduction	199
2.	Extending Probabilistic Models to Categorical Data	201
2.1	Modeling Mixed Data	203
3.	Extending Linear Models to Categorical and Mixed Data	204
4.	Extending Proximity Models to Categorical Data	205
4.1	Aggregate Statistical Similarity	206
4.2	Contextual Similarity	207
4.3	Issues with Mixed Data	209
4.4	Density-based Methods	210
4.5	Clustering Methods	210
5.	Outlier Detection in Binary and Transaction Data	210
5.1	Subspace Methods	211
5.2	Novelties in Temporal Transactions	212
6.	Outlier Detection in Text Data	213
6.1	Latent Semantic Indexing	213
6.2	First Story Detection	214
7.	Conclusions and Summary	220
8.	Bibliographic Survey	220
9.	Exercises	223

8

Time Series and Multidimensional Streaming Outlier Detection	225
1. Introduction	225
2. Prediction-based Outlier Detection of Streaming Time Series	229
2.1 Autoregressive Models	230
2.2 Multiple Time Series Regression Models	232
2.3 Supervised Outlier Detection in Time Series	237
3. Time-Series of Unusual Shapes	239
3.1 Transformation to Other Representations	241
3.2 Distance-based Methods	243
3.3 Single Series versus Multiple Series	245
3.4 Finding Unusual Shapes from Multivariate Series	246
3.5 Supervised Methods for Finding Unusual Time-Series Shapes	248
4. Outlier Detection in Multidimensional Data Streams	249
4.1 Individual Data Points as Outliers	250
4.2 Aggregate Change Points as Outliers	252
4.3 Rare and Novel Class Detection in Multidimensional Data Streams	257
5. Conclusions and Summary	260
6. Bibliographic Survey	260
7. Exercises	264

9

Outlier Detection in Discrete Sequences	267
1. Introduction	267
2. Position Outliers	270
2.1 Rule-based Models	273
2.2 Markovian Models	274
2.3 Efficiency Issues: Probabilistic Suffix Trees	277
3. Combination Outliers	280
3.1 A Primitive Model for Combination Outlier Detection	283
3.2 Distance-based Models	286
3.3 Frequency-based Models	290
3.4 Hidden Markov Models	292
4. Complex Sequences and Scenarios	304
4.1 Multivariate Sequences	304
4.2 Set-based Sequences	305
4.3 Online Applications: Early Anomaly Detection	306
5. Supervised Outliers in Sequences	306
6. Conclusions and Summary	309
7. Bibliographic Survey	309
8. Exercises	311

10

Spatial Outlier Detection	313
1. Introduction	313
2. Neighborhood-based Algorithms	318
2.1 Multidimensional Methods	319
2.2 Graph-based Methods	320
2.3 Handling Multiple Behavioral Attributes	321
3. Autoregressive Models	321
4. Visualization with Variogram Clouds	323

5.	Finding Abnormal Shapes in Spatial Data	326
6.	Spatio-temporal Outliers	332
6.1	Spatiotemporal Data: Trajectories	334
6.2	Anomalous Shape Change Detection	336
7.	Supervised Outlier Detection	336
7.1	Supervised Shape Discovery	336
7.2	Supervised Trajectory Discovery	338
8.	Conclusions and Summary	338
9.	Bibliographic Survey	339
10.	Exercises	341
11		
	Outlier Detection in Graphs and Networks	343
1.	Introduction	343
2.	Outlier Detection in Many Small Graphs	345
3.	Outlier Detection in a Single Large Graph	346
3.1	Node Outliers	347
3.2	Linkage Outliers	348
3.3	Subgraph Outliers	353
4.	Node Content in Outlier Analysis	354
5.	Change-based Outliers in Temporal Graphs	356
5.1	Stream Oriented Processing for Linkage Anomalies	357
5.2	Outliers based on Community Evolution	361
5.3	Outliers based on Shortest Path Distance Changes	367
5.4	Temporal Pattern-based Outliers	368
6.	Conclusions and Summary	368
7.	Bibliographic Survey	369
8.	Exercises	371
12		
	Applications of Outlier Analysis	373
1.	Introduction	373
2.	Quality Control and Fault Detection Applications	375
3.	Financial Applications	379
4.	Web Log Analytics	382
5.	Intrusion and Security Applications	384
6.	Medical Applications	387
7.	Text and Social Media Applications	389
8.	Earth Science Applications	391
9.	Miscellaneous Applications	394
10.	Guidelines for the Practitioner	396
11.	Resources for the Practitioner	398
12.	Conclusions and Summary	399
	References	401
	Index	443