
Contents

1	An Introduction to Outlier Analysis	1
1.1	Introduction	1
1.2	The Data Model is Everything	5
1.2.1	Connections with Supervised Models	8
1.3	The Basic Outlier Detection Models	10
1.3.1	Feature Selection in Outlier Detection	10
1.3.2	Extreme-Value Analysis	11
1.3.3	Probabilistic and Statistical Models	12
1.3.4	Linear Models	13
1.3.4.1	Spectral Models	14
1.3.5	Proximity-Based Models	14
1.3.6	Information-Theoretic Models	16
1.3.7	High-Dimensional Outlier Detection	17
1.4	Outlier Ensembles	18
1.4.1	Sequential Ensembles	19
1.4.2	Independent Ensembles	20
1.5	The Basic Data Types for Analysis	21
1.5.1	Categorical, Text, and Mixed Attributes	21
1.5.2	When the Data Values have Dependencies	21
1.5.2.1	Times-Series Data and Data Streams	22
1.5.2.2	Discrete Sequences	24
1.5.2.3	Spatial Data	24
1.5.2.4	Network and Graph Data	25
1.6	Supervised Outlier Detection	25
1.7	Outlier Evaluation Techniques	26
1.7.1	Interpreting the ROC AUC	29
1.7.2	Common Mistakes in Benchmarking	30
1.8	Conclusions and Summary	31
1.9	Bibliographic Survey	31
1.10	Exercises	33

2	Probabilistic Models for Outlier Detection	35
2.1	Introduction	35
2.2	Statistical Methods for Extreme-Value Analysis	37
2.2.1	Probabilistic Tail Inequalities	37
2.2.1.1	Sum of Bounded Random Variables	38
2.2.2	Statistical-Tail Confidence Tests	43
2.2.2.1	t -Value Test	43
2.2.2.2	Sum of Squares of Deviations	45
2.2.2.3	Visualizing Extreme Values with Box Plots	45
2.3	Extreme-Value Analysis in Multivariate Data	46
2.3.1	Depth-Based Methods	47
2.3.2	Deviation-Based Methods	48
2.3.3	Angle-Based Outlier Detection	49
2.3.4	Distance Distribution-based Techniques: The Mahalanobis Method	51
2.3.4.1	Strengths of the Mahalanobis Method	53
2.4	Probabilistic Mixture Modeling for Outlier Analysis	54
2.4.1	Relationship with Clustering Methods	57
2.4.2	The Special Case of a Single Mixture Component	58
2.4.3	Other Ways of Leveraging the EM Model	58
2.4.4	An Application of EM for Converting Scores to Probabilities	59
2.5	Limitations of Probabilistic Modeling	60
2.6	Conclusions and Summary	61
2.7	Bibliographic Survey	61
2.8	Exercises	62
3	Linear Models for Outlier Detection	65
3.1	Introduction	65
3.2	Linear Regression Models	68
3.2.1	Modeling with Dependent Variables	70
3.2.1.1	Applications of Dependent Variable Modeling	73
3.2.2	Linear Modeling with Mean-Squared Projection Error	74
3.3	Principal Component Analysis	75
3.3.1	Connections with the Mahalanobis Method	78
3.3.2	Hard PCA versus Soft PCA	79
3.3.3	Sensitivity to Noise	79
3.3.4	Normalization Issues	80
3.3.5	Regularization Issues	80
3.3.6	Applications to Noise Correction	80
3.3.7	How Many Eigenvectors?	81
3.3.8	Extension to Nonlinear Data Distributions	83
3.3.8.1	Choice of Similarity Matrix	85
3.3.8.2	Practical Issues	86
3.3.8.3	Application to Arbitrary Data Types	88
3.4	One-Class Support Vector Machines	88
3.4.1	Solving the Dual Optimization Problem	92
3.4.2	Practical Issues	92
3.4.3	Connections to Support Vector Data Description and Other Kernel Models	93
3.5	A Matrix Factorization View of Linear Models	95

3.5.1	Outlier Detection in Incomplete Data	96
3.5.1.1	Computing the Outlier Scores	98
3.6	Neural Networks: From Linear Models to Deep Learning	98
3.6.1	Generalization to Nonlinear Models	101
3.6.2	Replicator Neural Networks and Deep Autoencoders	102
3.6.3	Practical Issues	105
3.6.4	The Broad Potential of Neural Networks	106
3.7	Limitations of Linear Modeling	106
3.8	Conclusions and Summary	107
3.9	Bibliographic Survey	108
3.10	Exercises	109
4	Proximity-Based Outlier Detection	111
4.1	Introduction	111
4.2	Clusters and Outliers: The Complementary Relationship	112
4.2.1	Extensions to Arbitrarily Shaped Clusters	115
4.2.1.1	Application to Arbitrary Data Types	118
4.2.2	Advantages and Disadvantages of Clustering Methods	118
4.3	Distance-Based Outlier Analysis	118
4.3.1	Scoring Outputs for Distance-Based Methods	119
4.3.2	Binary Outputs for Distance-Based Methods	121
4.3.2.1	Cell-Based Pruning	122
4.3.2.2	Sampling-Based Pruning	124
4.3.2.3	Index-Based Pruning	126
4.3.3	Data-Dependent Similarity Measures	128
4.3.4	ODIN: A Reverse Nearest Neighbor Approach	129
4.3.5	Intensional Knowledge of Distance-Based Outliers	130
4.3.6	Discussion of Distance-Based Methods	131
4.4	Density-Based Outliers	131
4.4.1	LOF: Local Outlier Factor	132
4.4.1.1	Handling Duplicate Points and Stability Issues	134
4.4.2	LOCI: Local Correlation Integral	135
4.4.2.1	LOCI Plot	136
4.4.3	Histogram-Based Techniques	137
4.4.4	Kernel Density Estimation	138
4.4.4.1	Connection with Harmonic k -Nearest Neighbor Detector	139
4.4.4.2	Local Variations of Kernel Methods	140
4.4.5	Ensemble-Based Implementations of Histograms and Kernel Methods	140
4.5	Limitations of Proximity-Based Detection	141
4.6	Conclusions and Summary	142
4.7	Bibliographic Survey	142
4.8	Exercises	146
5	High-Dimensional Outlier Detection	149
5.1	Introduction	149
5.2	Axis-Parallel Subspaces	152
5.2.1	Genetic Algorithms for Outlier Detection	153
5.2.1.1	Defining Abnormal Lower-Dimensional Projections	153
5.2.1.2	Defining Genetic Operators for Subspace Search	154

5.2.2	Finding Distance-Based Outlying Subspaces	157
5.2.3	Feature Bagging: A Subspace Sampling Perspective	157
5.2.4	Projected Clustering Ensembles	158
5.2.5	Subspace Histograms in Linear Time	160
5.2.6	Isolation Forests	161
5.2.6.1	Further Enhancements for Subspace Selection	163
5.2.6.2	Early Termination	163
5.2.6.3	Relationship to Clustering Ensembles and Histograms	164
5.2.7	Selecting High-Contrast Subspaces	164
5.2.8	Local Selection of Subspace Projections	166
5.2.9	Distance-Based Reference Sets	169
5.3	Generalized Subspaces	170
5.3.1	Generalized Projected Clustering Approach	171
5.3.2	Leveraging Instance-Specific Reference Sets	172
5.3.3	Rotated Subspace Sampling	175
5.3.4	Nonlinear Subspaces	176
5.3.5	Regression Modeling Techniques	178
5.4	Discussion of Subspace Analysis	178
5.5	Conclusions and Summary	180
5.6	Bibliographic Survey	181
5.7	Exercises	184
6	Outlier Ensembles	185
6.1	Introduction	185
6.2	Categorization and Design of Ensemble Methods	188
6.2.1	Basic Score Normalization and Combination Methods	189
6.3	Theoretical Foundations of Outlier Ensembles	191
6.3.1	What is the Expectation Computed Over?	195
6.3.2	Relationship of Ensemble Analysis to Bias-Variance Trade-Off	195
6.4	Variance Reduction Methods	196
6.4.1	Parametric Ensembles	197
6.4.2	Randomized Detector Averaging	199
6.4.3	Feature Bagging: An Ensemble-Centric Perspective	199
6.4.3.1	Connections to Representational Bias	200
6.4.3.2	Weaknesses of Feature Bagging	202
6.4.4	Rotated Bagging	202
6.4.5	Isolation Forests: An Ensemble-Centric View	203
6.4.6	Data-Centric Variance Reduction with Sampling	205
6.4.6.1	Bagging	205
6.4.6.2	Subsampling	206
6.4.6.3	Variable Subsampling	207
6.4.6.4	Variable Subsampling with Rotated Bagging (VR)	209
6.4.7	Other Variance Reduction Methods	209
6.5	Flying Blind with Bias Reduction	211
6.5.1	Bias Reduction by Data-Centric Pruning	211
6.5.2	Bias Reduction by Model-Centric Pruning	212
6.5.3	Combining Bias and Variance Reduction	213
6.6	Model Combination for Outlier Ensembles	214
6.6.1	Combining Scoring Methods with Ranks	215

6.6.2	Combining Bias and Variance Reduction	216
6.7	Conclusions and Summary	217
6.8	Bibliographic Survey	217
6.9	Exercises	218
7	Supervised Outlier Detection	219
7.1	Introduction	219
7.2	Full Supervision: Rare Class Detection	221
7.2.1	Cost-Sensitive Learning	223
7.2.1.1	MetaCost: A Relabeling Approach	223
7.2.1.2	Weighting Methods	225
7.2.2	Adaptive Re-sampling	228
7.2.2.1	Relationship between Weighting and Sampling	229
7.2.2.2	Synthetic Over-sampling: SMOTE	229
7.2.3	Boosting Methods	230
7.3	Semi-Supervision: Positive and Unlabeled Data	231
7.4	Semi-Supervision: Partially Observed Classes	232
7.4.1	One-Class Learning with Anomalous Examples	233
7.4.2	One-Class Learning with Normal Examples	234
7.4.3	Learning with a Subset of Labeled Classes	234
7.5	Unsupervised Feature Engineering in Supervised Methods	235
7.6	Active Learning	236
7.7	Supervised Models for Unsupervised Outlier Detection	239
7.7.1	Connections with PCA-Based Methods	242
7.7.2	Group-wise Predictions for High-Dimensional Data	243
7.7.3	Applicability to Mixed-Attribute Data Sets	244
7.7.4	Incorporating Column-wise Knowledge	244
7.7.5	Other Classification Methods with Synthetic Outliers	244
7.8	Conclusions and Summary	245
7.9	Bibliographic Survey	245
7.10	Exercises	247
8	Categorical, Text, and Mixed Attribute Data	249
8.1	Introduction	249
8.2	Extending Probabilistic Models to Categorical Data	250
8.2.1	Modeling Mixed Data	253
8.3	Extending Linear Models to Categorical and Mixed Data	254
8.3.1	Leveraging Supervised Regression Models	254
8.4	Extending Proximity Models to Categorical Data	255
8.4.1	Aggregate Statistical Similarity	256
8.4.2	Contextual Similarity	257
8.4.2.1	Connections to Linear Models	258
8.4.3	Issues with Mixed Data	259
8.4.4	Density-Based Methods	259
8.4.5	Clustering Methods	259
8.5	Outlier Detection in Binary and Transaction Data	260
8.5.1	Subspace Methods	260
8.5.2	Novelties in Temporal Transactions	262
8.6	Outlier Detection in Text Data	262

8.6.1	Probabilistic Models	262
8.6.2	Linear Models: Latent Semantic Analysis	264
8.6.2.1	Probabilistic Latent Semantic Analysis (PLSA)	265
8.6.3	Proximity-Based Models	268
8.6.3.1	First Story Detection	269
8.7	Conclusions and Summary	270
8.8	Bibliographic Survey	270
8.9	Exercises	272
9	Time Series and Streaming Outlier Detection	273
9.1	Introduction	273
9.2	Predictive Outlier Detection in Streaming Time-Series	276
9.2.1	Autoregressive Models	276
9.2.2	Multiple Time Series Regression Models	279
9.2.2.1	Direct Generalization of Autoregressive Models	279
9.2.2.2	Time-Series Selection Methods	281
9.2.2.3	Principal Component Analysis and Hidden Variable-Based Models	282
9.2.3	Relationship between Unsupervised Outlier Detection and Prediction	284
9.2.4	Supervised Point Outlier Detection in Time Series	284
9.3	Time-Series of Unusual Shapes	286
9.3.1	Transformation to Other Representations	287
9.3.1.1	Numeric Multidimensional Transformations	288
9.3.1.2	Discrete Sequence Transformations	290
9.3.1.3	Leveraging Trajectory Representations of Time Series	291
9.3.2	Distance-Based Methods	293
9.3.2.1	Single Series versus Multiple Series	295
9.3.3	Probabilistic Models	295
9.3.4	Linear Models	295
9.3.4.1	Univariate Series	295
9.3.4.2	Multivariate Series	296
9.3.4.3	Incorporating Arbitrary Similarity Functions	297
9.3.4.4	Leveraging Kernel Methods with Linear Models	298
9.3.5	Supervised Methods for Finding Unusual Time-Series Shapes	298
9.4	Multidimensional Streaming Outlier Detection	298
9.4.1	Individual Data Points as Outliers	299
9.4.1.1	Proximity-Based Algorithms	299
9.4.1.2	Probabilistic Algorithms	301
9.4.1.3	High-Dimensional Scenario	301
9.4.2	Aggregate Change Points as Outliers	301
9.4.2.1	Velocity Density Estimation Method	302
9.4.2.2	Statistically Significant Changes in Aggregate Distributions	304
9.4.3	Rare and Novel Class Detection in Multidimensional Data Streams	305
9.4.3.1	Detecting Rare Classes	305
9.4.3.2	Detecting Novel Classes	306
9.4.3.3	Detecting Infrequently Recurring Classes	306
9.5	Conclusions and Summary	307
9.6	Bibliographic Survey	307
9.7	Exercises	310

10 Outlier Detection in Discrete Sequences	311
10.1 Introduction	311
10.2 Position Outliers	313
10.2.1 Rule-Based Models	315
10.2.2 Markovian Models	316
10.2.3 Efficiency Issues: Probabilistic Suffix Trees	318
10.3 Combination Outliers	320
10.3.1 A Primitive Model for Combination Outlier Detection	322
10.3.1.1 Model-Specific Combination Issues	323
10.3.1.2 Easier Special Cases	323
10.3.1.3 Relationship between Position and Combination Outliers	324
10.3.2 Distance-Based Models	324
10.3.2.1 Combining Anomaly Scores from Comparison Units	326
10.3.2.2 Some Observations on Distance-Based Methods	327
10.3.2.3 Easier Special Case: Short Sequences	327
10.3.3 Frequency-Based Models	327
10.3.3.1 Frequency-Based Model with User-Specified Comparison Unit	327
10.3.3.2 Frequency-Based Model with Extracted Comparison Units	328
10.3.3.3 Combining Anomaly Scores from Comparison Units	329
10.3.4 Hidden Markov Models	329
10.3.4.1 Design Choices in a Hidden Markov Model	331
10.3.4.2 Training and Prediction with HMMs	333
10.3.4.3 Evaluation: Computing the Fit Probability for Observed Se-	
quences	334
10.3.4.4 Explanation: Determining the Most Likely State Sequence	
for Observed Sequence	334
10.3.4.5 Training: Baum-Welch Algorithm	335
10.3.4.6 Computing Anomaly Scores	336
10.3.4.7 Special Case: Short Sequence Anomaly Detection	337
10.3.5 Kernel-Based Methods	337
10.4 Complex Sequences and Scenarios	338
10.4.1 Multivariate Sequences	338
10.4.2 Set-Based Sequences	339
10.4.3 Online Applications: Early Anomaly Detection	340
10.5 Supervised Outliers in Sequences	340
10.6 Conclusions and Summary	342
10.7 Bibliographic Survey	342
10.8 Exercises	344
11 Spatial Outlier Detection	345
11.1 Introduction	345
11.2 Spatial Attributes are Contextual	349
11.2.1 Neighborhood-Based Algorithms	349
11.2.1.1 Multidimensional Methods	350
11.2.1.2 Graph-Based Methods	351
11.2.1.3 The Case of Multiple Behavioral Attributes	351
11.2.2 Autoregressive Models	352
11.2.3 Visualization with Variogram Clouds	353
11.2.4 Finding Abnormal Shapes in Spatial Data	355

11.2.4.1	Contour Extraction Methods	356
11.2.4.2	Extracting Multidimensional Representations	360
11.2.4.3	Multidimensional Wavelet Transformation	360
11.2.4.4	Supervised Shape Discovery	360
11.2.4.5	Anomalous Shape Change Detection	361
11.3	Spatiotemporal Outliers with Spatial and Temporal Context	362
11.4	Spatial Behavior with Temporal Context: Trajectories	363
11.4.1	Real-Time Anomaly Detection	363
11.4.2	Unusual Trajectory Shapes	363
11.4.2.1	Segment-wise Partitioning Methods	363
11.4.2.2	Tile-Based Transformations	364
11.4.2.3	Similarity-Based Transformations	365
11.4.3	Supervised Outliers in Trajectories	365
11.5	Conclusions and Summary	366
11.6	Bibliographic Survey	366
11.7	Exercises	367
12	Outlier Detection in Graphs and Networks	369
12.1	Introduction	369
12.2	Outlier Detection in Many Small Graphs	371
12.2.1	Leveraging Graph Kernels	371
12.3	Outlier Detection in a Single Large Graph	372
12.3.1	Node Outliers	372
12.3.1.1	Leveraging the Mahalanobis Method	374
12.3.2	Linkage Outliers	374
12.3.2.1	Matrix Factorization Methods	374
12.3.2.2	Spectral Methods and Embeddings	378
12.3.2.3	Clustering Methods	379
12.3.2.4	Community Linkage Outliers	380
12.3.3	Subgraph Outliers	381
12.4	Node Content in Outlier Analysis	382
12.4.1	Shared Matrix Factorization	382
12.4.2	Relating Feature Similarity to Tie Strength	383
12.4.3	Heterogeneous Markov Random Fields	384
12.5	Change-Based Outliers in Temporal Graphs	384
12.5.1	Discovering Node Hotspots in Graph Streams	385
12.5.2	Streaming Detection of Linkage Anomalies	386
12.5.3	Outliers Based on Community Evolution	388
12.5.3.1	Integrating Clustering Maintenance with Evolution Analysis	388
12.5.3.2	Online Analysis of Community Evolution in Graph Streams	390
12.5.3.3	GraphScope	390
12.5.4	Outliers Based on Shortest Path Distance Changes	392
12.5.5	Matrix Factorization and Latent Embedding Methods	392
12.6	Conclusions and Summary	393
12.7	Bibliographic Survey	394
12.8	Exercises	396

13 Applications of Outlier Analysis	399
13.1 Introduction	399
13.2 Quality Control and Fault Detection Applications	401
13.3 Financial Applications	404
13.4 Web Log Analytics	406
13.5 Intrusion and Security Applications	407
13.6 Medical Applications	410
13.7 Text and Social Media Applications	411
13.8 Earth Science Applications	413
13.9 Miscellaneous Applications	415
13.10 Guidelines for the Practitioner	416
13.10.1 Which Unsupervised Algorithms Work Best?	418
13.11 Resources for the Practitioner	421
13.12 Conclusions and Summary	422

Outlier Analysis

Aggarwal, C.C.

2017, XXII, 466 p. 78 illus., 13 illus. in color., Hardcover

ISBN: 978-3-319-47577-6