Outlier Analysis
To my wife, my daughter Sayani,
and my late parents Dr. Prem Sarup and Mrs. Pushplata Aggarwal.
## Contents

1 **An Introduction to Outlier Analysis**  
   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

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Introduction</td>
<td>35</td>
</tr>
<tr>
<td>2.2</td>
<td>Statistical Methods for Extreme-Value Analysis</td>
<td>37</td>
</tr>
<tr>
<td>2.2.1</td>
<td>Probabilistic Tail Inequalities</td>
<td>37</td>
</tr>
<tr>
<td>2.2.1.1</td>
<td>Sum of Bounded Random Variables</td>
<td>38</td>
</tr>
<tr>
<td>2.2.2</td>
<td>Statistical-Tail Confidence Tests</td>
<td>43</td>
</tr>
<tr>
<td>2.2.2.1</td>
<td>$t$-Value Test</td>
<td>43</td>
</tr>
<tr>
<td>2.2.2.2</td>
<td>Sum of Squares of Deviations</td>
<td>45</td>
</tr>
<tr>
<td>2.2.2.3</td>
<td>Visualizing Extreme Values with Box Plots</td>
<td>45</td>
</tr>
<tr>
<td>2.3</td>
<td>Extreme-Value Analysis in Multivariate Data</td>
<td>46</td>
</tr>
<tr>
<td>2.3.1</td>
<td>Depth-Based Methods</td>
<td>47</td>
</tr>
<tr>
<td>2.3.2</td>
<td>Deviation-Based Methods</td>
<td>48</td>
</tr>
<tr>
<td>2.3.3</td>
<td>Angle-Based Outlier Detection</td>
<td>49</td>
</tr>
<tr>
<td>2.3.4</td>
<td>Distance Distribution-based Techniques: The Mahalanobis Method</td>
<td>51</td>
</tr>
<tr>
<td>2.3.4.1</td>
<td>Strengths of the Mahalanobis Method</td>
<td>53</td>
</tr>
<tr>
<td>2.4</td>
<td>Probabilistic Mixture Modeling for Outlier Analysis</td>
<td>54</td>
</tr>
<tr>
<td>2.4.1</td>
<td>Relationship with Clustering Methods</td>
<td>57</td>
</tr>
<tr>
<td>2.4.2</td>
<td>The Special Case of a Single Mixture Component</td>
<td>58</td>
</tr>
<tr>
<td>2.4.3</td>
<td>Other Ways of Leveraging the EM Model</td>
<td>58</td>
</tr>
<tr>
<td>2.4.4</td>
<td>An Application of EM for Converting Scores to Probabilities</td>
<td>59</td>
</tr>
<tr>
<td>2.5</td>
<td>Limitations of Probabilistic Modeling</td>
<td>60</td>
</tr>
<tr>
<td>2.6</td>
<td>Conclusions and Summary</td>
<td>61</td>
</tr>
<tr>
<td>2.7</td>
<td>Bibliographic Survey</td>
<td>61</td>
</tr>
<tr>
<td>2.8</td>
<td>Exercises</td>
<td>62</td>
</tr>
</tbody>
</table>

## 3 Linear Models for Outlier Detection

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Introduction</td>
<td>65</td>
</tr>
<tr>
<td>3.2</td>
<td>Linear Regression Models</td>
<td>68</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Modeling with Dependent Variables</td>
<td>70</td>
</tr>
<tr>
<td>3.2.1.1</td>
<td>Applications of Dependent Variable Modeling</td>
<td>73</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Linear Modeling with Mean-Squared Projection Error</td>
<td>74</td>
</tr>
<tr>
<td>3.3</td>
<td>Principal Component Analysis</td>
<td>75</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Connections with the Mahalanobis Method</td>
<td>78</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Hard PCA versus Soft PCA</td>
<td>79</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Sensitivity to Noise</td>
<td>79</td>
</tr>
<tr>
<td>3.3.4</td>
<td>Normalization Issues</td>
<td>80</td>
</tr>
<tr>
<td>3.3.5</td>
<td>Regularization Issues</td>
<td>80</td>
</tr>
<tr>
<td>3.3.6</td>
<td>Applications to Noise Correction</td>
<td>80</td>
</tr>
<tr>
<td>3.3.7</td>
<td>How Many Eigenvectors?</td>
<td>81</td>
</tr>
<tr>
<td>3.3.8</td>
<td>Extension to Nonlinear Data Distributions</td>
<td>83</td>
</tr>
<tr>
<td>3.3.8.1</td>
<td>Choice of Similarity Matrix</td>
<td>85</td>
</tr>
<tr>
<td>3.3.8.2</td>
<td>Practical Issues</td>
<td>86</td>
</tr>
<tr>
<td>3.3.8.3</td>
<td>Application to Arbitrary Data Types</td>
<td>88</td>
</tr>
<tr>
<td>3.4</td>
<td>One-Class Support Vector Machines</td>
<td>88</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Solving the Dual Optimization Problem</td>
<td>92</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Practical Issues</td>
<td>92</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Connections to Support Vector Data Description and Other Kernel Models</td>
<td>93</td>
</tr>
<tr>
<td>3.5</td>
<td>A Matrix Factorization View of Linear Models</td>
<td>95</td>
</tr>
</tbody>
</table>
CONTENTS

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
## 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
- Page 262

### 8.6.2 Linear Models: Latent Semantic Analysis
  - 8.6.2.1 Probabilistic Latent Semantic Analysis (PLSA) - Page 265

### 8.7 Conclusions and Summary
- Page 270

### 8.8 Bibliographic Survey
- Page 270

### 8.9 Exercises
- Page 272

### 9 Time Series and Streaming Outlier Detection
- Page 273

#### 9.1 Introduction
- Page 273

#### 9.2 Predictive Outlier Detection in Streaming Time-Series
  - 9.2.1 Autoregressive Models - Page 276
  - 9.2.2 Multiple Time Series Regression Models - Page 279
    - 9.2.2.1 Direct Generalization of Autoregressive Models - Page 279
    - 9.2.2.2 Time-Series Selection Methods - Page 281
    - 9.2.2.3 Principal Component Analysis and Hidden Variable-Based Models - Page 282
  - 9.2.3 Relationship between Unsupervised Outlier Detection and Prediction - Page 284
  - 9.2.4 Supervised Point Outlier Detection in Time Series - Page 284

#### 9.3 Time-Series of Unusual Shapes
  - Page 286
    - 9.3.1 Transformation to Other Representations
      - 9.3.1.1 Numeric Multidimensional Transformations - Page 288
      - 9.3.1.2 Discrete Sequence Transformations - Page 290
      - 9.3.1.3 Leveraging Trajectory Representations of Time Series - Page 291
  - 9.3.2 Distance-Based Methods - Page 293
    - 9.3.2.1 Single Series versus Multiple Series - Page 295
  - 9.3.3 Probabilistic Models - Page 295
  - 9.3.4 Linear Models - Page 295
    - 9.3.4.1 Univariate Series - Page 295
    - 9.3.4.2 Multivariate Series - Page 296
    - 9.3.4.3 Incorporating Arbitrary Similarity Functions - Page 297
    - 9.3.4.4 Leveraging Kernel Methods with Linear Models - Page 298
  - 9.3.5 Supervised Methods for Finding Unusual Time-Series Shapes - Page 298

#### 9.4 Multidimensional Streaming Outlier Detection
  - Page 298
    - 9.4.1 Individual Data Points as Outliers
      - 9.4.1.1 Proximity-Based Algorithms - Page 299
      - 9.4.1.2 Probabilistic Algorithms - Page 301
      - 9.4.1.3 High-Dimensional Scenario - Page 301
    - 9.4.2 Aggregate Change Points as Outliers
      - 9.4.2.1 Velocity Density Estimation Method - Page 302
      - 9.4.2.2 Statistically Significant Changes in Aggregate Distributions - Page 304
    - 9.4.3 Rare and Novel Class Detection in Multidimensional Data Streams - Page 305
      - 9.4.3.1 Detecting Rare Classes - Page 305
      - 9.4.3.2 Detecting Novel Classes - Page 306
      - 9.4.3.3 Detecting Infrequently Recurring Classes - Page 306

#### 9.5 Conclusions and Summary
- Page 307

#### 9.6 Bibliographic Survey
- Page 307

#### 9.7 Exercises
- Page 310
10 Outlier Detection in Discrete Sequences

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 Units ... 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 Sequences .... 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
## Contents

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 Spatialtemporal 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

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

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
Preface

“All things excellent are as difficult as they are rare.” – Baruch Spinoza

First Edition

Most of the earliest work on outlier detection was performed by the statistics community. While statistical methods are mathematically more precise, they have several shortcomings, such as simplified assumptions about data representations, poor algorithmic scalability, and a low focus on interpretability. With the increasing advances in hardware technology for data collection, and advances in software technology (databases) for data organization, computer scientists have increasingly been participating in the latest advancements of this field. Computer scientists approach this field based on their practical experiences in managing large amounts of data, and with far fewer assumptions – the data can be of any type, structured or unstructured, and may be extremely large. Furthermore, issues such as computational efficiency and intuitive analysis of the data are generally considered more important by computer scientists than mathematical precision, though the latter is important as well. This is the approach of professionals from the field of data mining, an area of computer science that was founded about 20 years ago. This has led to the formation of multiple academic communities on the subject, which have remained separated, partially because of differences in technical style and opinions about the importance of different problems and approaches to the subject. At this point, data mining professionals (with a computer science background) are much more actively involved in this area as compared to statisticians. This seems to be a major change in the research landscape. This book presents outlier detection from an integrated perspective, though the focus is towards computer science professionals. Special emphasis was placed on relating the methods from different communities with one another.

The key advantage of writing the book at this point in time is that the vast amount of work done by computer science professionals in the last two decades has remained largely untouched by a formal book on the subject. The classical books relevant to outlier analysis are as follows:


xvii
We note that these books are quite outdated, and the most recent among them is a decade old. Furthermore, this (most recent) book is really focused on the relationship between regression and outlier analysis, rather than the latter. Outlier analysis is a much broader area, in which regression analysis is only a small part. The other books are even older, and are between 15 and 25 years old. They are exclusively targeted to the statistics community. This is not surprising, given that the first mainstream computer science conference in data mining (KDD) was organized in 1995. Most of the work in the data-mining community was performed after the writing of these books. Therefore, many key topics of interest to the broader data mining community are not covered in these books. Given that outlier analysis has been explored by a much broader community, including databases, data mining, statistics, and machine learning, we feel that our book incorporates perspectives from a much broader audience and brings together different points of view.

The chapters of this book have been organized carefully, with a view of covering the area extensively in a natural order. Emphasis was placed on simplifying the content, so that students and practitioners can also benefit from the book. While we did not originally intend to create a textbook on the subject, it evolved during the writing process into a work that can also be used as a teaching aid. Furthermore, it can also be used as a reference book, since each chapter contains extensive bibliographic notes. Therefore, this book serves a dual purpose, providing a comprehensive exposition of the topic of outlier detection from multiple points of view.

**Additional Notes for the Second Edition**

The second edition of this book is a significant enhancement over the first edition. In particular, most of the chapters have been upgraded with new material and recent techniques. More explanations have been added at several places and newer techniques have also been added. An entire chapter on outlier ensembles has been added. Many new topics have been added to the book such as feature selection, one-class support vector machines, one-class neural networks, matrix factorization, spectral methods, wavelet transforms, and supervised learning. Every chapter has been updated with the latest algorithms on the topic.

Last but not least, the first edition was classified by the publisher as a monograph, whereas the second edition is formally classified as a textbook. The writing style has been enhanced to be easily understandable to students. Many algorithms have been described in greater detail, as one might expect from a textbook. It is also accompanied with a solution manual for classroom teaching.
Acknowledgments

First Edition

I would like to thank my wife and daughter for their love and support during the writing of this book. The writing of a book requires significant time that is taken away from family members. This book is the result of their patience with me during this time. I also owe my late parents a debt of gratitude for instilling in me a love of education, which has played an important inspirational role in my book-writing efforts.

I would also like to thank my manager Nagui Halim for providing the tremendous support necessary for the writing of this book. His professional support has been instrumental for my many book efforts in the past and present.

Over the years, I have benefited from the insights of numerous collaborators. An incomplete list of these long-term collaborators in alphabetical order is Tarek F. Abdelzaher, Jiawei Han, Thomas S. Huang, Latifur Khan, Mohammad M. Masud, Spiros Papadimitriou, Guojun Qi, and Philip S. Yu. I would like to thank them for their collaborations and insights over the course of many years.

I would also like to specially thank my advisor James B. Orlin for his guidance during my early years as a researcher. While I no longer work in the same area, the legacy of what I learned from him is a crucial part of my approach to research. In particular, he taught me the importance of intuition and simplicity of thought in the research process. These are more important aspects of research than is generally recognized. This book is written in a simple and intuitive style, and is meant to improve accessibility of this area to both researchers and practitioners.

Finally, I would like to thank Lata Aggarwal for helping me with some of the figures created using PowerPoint graphics in this book.

Acknowledgments for Second Edition

I received significant feedback from various colleagues during the writing of the second edition. In particular, I would like to acknowledge Leman Akoglu, Chih-Jen Lin, Saket Sathe, Jiliang Tang, and Suhang Wang. Leman and Saket provided detailed feedback on several sections and chapters of this book.
Charu C. Aggarwal is a Distinguished Research Staff Member (DRSM) at the IBM T. J. Watson Research Center in Yorktown Heights, New York. He completed his undergraduate degree in Computer Science from the Indian Institute of Technology at Kanpur in 1993 and his Ph.D. from the Massachusetts Institute of Technology in 1996. He has worked extensively in the field of data mining. He has published more than 300 papers in refereed conferences and journals and authored over 80 patents. He is the author or editor of 15 books, including a textbook on data mining and a comprehensive book on outlier analysis. Because of the commercial value of his patents, he has thrice been designated a Master Inventor at IBM. He is a recipient of an IBM Corporate Award (2003) for his work on bio-terrorist threat detection in data streams, a recipient of the IBM Outstanding Innovation Award (2008) for his scientific contributions to privacy technology, a recipient of two IBM Outstanding Technical Achievement Awards (2009, 2015) for his work on data streams and high-dimensional data, respectively. He received the EDBT 2014 Test of Time Award for his work on condensation-based privacy-preserving data mining. He is also a recipient of the IEEE ICDM Research Contributions Award (2015), which is one of the two highest awards for influential research contributions in the field of data mining.

He has served as the general co-chair of the IEEE Big Data Conference (2014) and as the program co-chair of the ACM CIKM Conference (2015), the IEEE ICDM Conference (2015), and the ACM KDD Conference (2016). He served as an associate editor of the IEEE Transactions on Knowledge and Data Engineering from 2004 to 2008. He is an associate editor of the ACM Transactions on Knowledge Discovery from Data, an associate editor of the IEEE Transactions on Big Data, an action editor of the Data Mining and Knowledge Discovery Journal, editor-in-chief of the ACM SIGKDD Explorations, and an associate editor of the Knowledge and Information Systems Journal. He serves on the advisory board of the Lecture Notes on Social Networks, a publication by Springer. He has served as the vice-president of the SIAM Activity Group on Data Mining and is a member of the SIAM industry committee. He is a fellow of the SIAM, ACM, and the IEEE, for “contributions to knowledge discovery and data mining algorithms.”