



INTRODUCTION TO DATA MINING





INTRODUCTION TO DATA MINING

SECOND EDITION

PANG-NING TAN
Michigan State University

MICHAEL STEINBACH
University of Minnesota

ANUJ KARPATNE
University of Minnesota

VIPIN KUMAR
University of Minnesota



330 Hudson Street, NY NY 10013



Director, Portfolio Management: Engineering, Computer Science & Global Editions: Julian Partridge
Specialist, Higher Ed Portfolio Management: Matt Goldstein
Portfolio Management Assistant: Meghan Jacoby
Managing Content Producer: Scott Disanno
Content Producer: Carole Snyder
Web Developer: Steve Wright
Rights and Permissions Manager: Ben Ferrini
Manufacturing Buyer, Higher Ed, Lake Side Communications Inc (LSC): Maura Zaldivar-Garcia
Inventory Manager: Ann Lam
Product Marketing Manager: Yvonne Vannatta
Field Marketing Manager: Demetrius Hall
Marketing Assistant: Jon Bryant
Cover Designer: Joyce Wells, jWellsDesign
Full-Service Project Management: Chandrasekar Subramanian, SPi Global
Composition: SPi Global

Copyright ©2019 Pearson Education, Inc. All rights reserved. Manufactured in the United States of America. This publication is protected by Copyright, and permission should be obtained from the publisher prior to any prohibited reproduction, storage in a retrieval system, or transmission in any form or by any means, electronic, mechanical, photocopying, recording, or likewise. For information regarding permissions, request forms and the appropriate contacts within the Pearson Education Global Rights & Permissions department, please visit www.pearsonhighed.com/permissions/.

Many of the designations by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and the publisher was aware of a trademark claim, the designations have been printed in initial caps or all caps.

The author and publisher of this book have used their best efforts in preparing this book. These efforts include the development, research, and testing of theories and programs to determine their effectiveness. The author and publisher make no warranty of any kind, expressed or implied, with regard to these programs or the documentation contained in this book. The author and publisher shall not be liable in any event for incidental or consequential damages with, or arising out of, the furnishing, performance or use of these programs.

Library of Congress Cataloging-in-Publication Data on File

Names: Tan, Pang-Ning, author. | Steinbach, Michael, author. | Karpatne, Anuj, author. | Kumar, Vipin, 1956- author.

Title: *Introduction to Data Mining* / Pang-Ning Tan, Michigan State University, Michael Steinbach, University of Minnesota, Anuj Karpatne, University of Minnesota, Vipin Kumar, University of Minnesota.

Description: Second edition. | New York, NY : Pearson Education, [2019] | Includes bibliographical references and index.

Identifiers: LCCN 2017048641 | ISBN 9780133128901 | ISBN 0133128903

Subjects: LCSH: Data mining.

Classification: LCC QA76.9.D343 T35 2019 | DDC 006.3/12–dc23 LC record available at <https://lcn.loc.gov/2017048641>



1 18
ISBN-10: 0-13-312890-3
ISBN-13: 978-0-13-312890-1



To our families ...





Preface to the Second Edition

Since the first edition, roughly 12 years ago, much has changed in the field of data analysis. The volume and variety of data being collected continues to increase, as has the rate (velocity) at which it is being collected and used to make decisions. Indeed, the term, Big Data, has been used to refer to the massive and diverse data sets now available. In addition, the term data science has been coined to describe an emerging area that applies tools and techniques from various fields, such as data mining, machine learning, statistics, and many others, to extract actionable insights from data, often big data.

The growth in data has created numerous opportunities for all areas of data analysis. The most dramatic developments have been in the area of predictive modeling, across a wide range of application domains. For instance, recent advances in neural networks, known as deep learning, have shown impressive results in a number of challenging areas, such as image classification, speech recognition, as well as text categorization and understanding. While not as dramatic, other areas, e.g., clustering, association analysis, and anomaly detection have also continued to advance. This new edition is in response to those advances.

Overview As with the first edition, the second edition of the book provides a comprehensive introduction to data mining and is designed to be accessible and useful to students, instructors, researchers, and professionals. Areas covered include data preprocessing, predictive modeling, association analysis, cluster analysis, anomaly detection, and avoiding false discoveries. The goal is to present fundamental concepts and algorithms for each topic, thus providing the reader with the necessary background for the application of data mining to real problems. As before, classification, association analysis and cluster analysis, are each covered in a pair of chapters. The introductory chapter covers basic concepts, representative algorithms, and evaluation techniques, while the more following chapter discusses advanced concepts and algorithms. As before, our objective is to provide the reader with a sound understanding of the foundations of data mining, while still covering many important advanced





vi Preface to the Second Edition

topics. Because of this approach, the book is useful both as a learning tool and as a reference.

To help readers better understand the concepts that have been presented, we provide an extensive set of examples, figures, and exercises. The solutions to the original exercises, which are already circulating on the web, will be made public. The exercises are mostly unchanged from the last edition, with the exception of new exercises in the chapter on avoiding false discoveries. New exercises for the other chapters and their solutions will be available to instructors via the web. Bibliographic notes are included at the end of each chapter for readers who are interested in more advanced topics, historically important papers, and recent trends. These have also been significantly updated. The book also contains a comprehensive subject and author index.

What is New in the Second Edition? Some of the most significant improvements in the text have been in the two chapters on classification. The introductory chapter uses the decision tree classifier for illustration, but the discussion on many topics—those that apply across all classification approaches—has been greatly expanded and clarified, including topics such as overfitting, underfitting, the impact of training size, model complexity, model selection, and common pitfalls in model evaluation. Almost every section of the advanced classification chapter has been significantly updated. The material on Bayesian networks, support vector machines, and artificial neural networks has been significantly expanded. We have added a separate section on deep networks to address the current developments in this area. The discussion of evaluation, which occurs in the section on imbalanced classes, has also been updated and improved.

The changes in association analysis are more localized. We have completely reworked the section on the evaluation of association patterns (introductory chapter), as well as the sections on sequence and graph mining (advanced chapter). Changes to cluster analysis are also localized. The introductory chapter added the K-means initialization technique and an updated the discussion of cluster evaluation. The advanced clustering chapter adds a new section on spectral graph clustering. Anomaly detection has been greatly revised and expanded. Existing approaches—statistical, nearest neighbor/density-based, and clustering based—have been retained and updated, while new approaches have been added: reconstruction-based, one-class classification, and information-theoretic. The reconstruction-based approach is illustrated using autoencoder networks that are part of the deep learning paradigm. The data chapter has



been updated to include discussions of mutual information and kernel-based techniques.

The last chapter, which discusses how to avoid false discoveries and produce valid results, is completely new, and is novel among other contemporary textbooks on data mining. It supplements the discussions in the other chapters with a discussion of the statistical concepts (statistical significance, p-values, false discovery rate, permutation testing, etc.) relevant to avoiding spurious results, and then illustrates these concepts in the context of data mining techniques. This chapter addresses the increasing concern over the validity and reproducibility of results obtained from data analysis. The addition of this last chapter is a recognition of the importance of this topic and an acknowledgment that a deeper understanding of this area is needed for those analyzing data.

The data exploration chapter has been deleted, as have the appendices, from the print edition of the book, but will remain available on the web. A new appendix provides a brief discussion of scalability in the context of big data.

To the Instructor As a textbook, this book is suitable for a wide range of students at the advanced undergraduate or graduate level. Since students come to this subject with diverse backgrounds that may not include extensive knowledge of statistics or databases, our book requires minimal prerequisites. No database knowledge is needed, and we assume only a modest background in statistics or mathematics, although such a background will make for easier going in some sections. As before, the book, and more specifically, the chapters covering major data mining topics, are designed to be as self-contained as possible. Thus, the order in which topics can be covered is quite flexible. The core material is covered in chapters 2 (data), 3 (classification), 5 (association analysis), 7 (clustering), and 9 (anomaly detection). We recommend at least a cursory coverage of Chapter 10 (Avoiding False Discoveries) to instill in students some caution when interpreting the results of their data analysis. Although the introductory data chapter (2) should be covered first, the basic classification (3), association analysis (5), and clustering chapters (7), can be covered in any order. Because of the relationship of anomaly detection (9) to classification (3) and clustering (7), these chapters should precede Chapter 9. Various topics can be selected from the advanced classification, association analysis, and clustering chapters (4, 6, and 8, respectively) to fit the schedule and interests of the instructor and students. We also advise that the lectures be augmented by projects or practical exercises in data mining. Although they



viii Preface to the Second Edition

are time consuming, such hands-on assignments greatly enhance the value of the course.

Support Materials Support materials available to all readers of this book are available at <http://www-users.cs.umn.edu/~kumar/dmbook/>.

- PowerPoint lecture slides
- Suggestions for student projects
- Data mining resources, such as algorithms and data sets
- Online tutorials that give step-by-step examples for selected data mining techniques described in the book using actual data sets and data analysis software

Additional support materials, including solutions to exercises, are available only to instructors adopting this textbook for classroom use. The book's resources will be mirrored at www.pearsonhighered.com/cs-resources. Comments and suggestions, as well as reports of errors, can be sent to the authors through dmbook@cs.umn.edu.

Acknowledgments Many people contributed to the first and second editions of the book. We begin by acknowledging our families to whom this book is dedicated. Without their patience and support, this project would have been impossible.

We would like to thank the current and former students of our data mining groups at the University of Minnesota and Michigan State for their contributions. Eui-Hong (Sam) Han and Mahesh Joshi helped with the initial data mining classes. Some of the exercises and presentation slides that they created can be found in the book and its accompanying slides. Students in our data mining groups who provided comments on drafts of the book or who contributed in other ways include Shyam Boriah, Haibin Cheng, Varun Chandola, Eric Eilertson, Levent Ertöz, Jing Gao, Rohit Gupta, Sridhar Iyer, Jung-Eun Lee, Benjamin Mayer, Aysel Ozgur, Uygur Oztekin, Gaurav Pandey, Kashif Riaz, Jerry Scripps, Gyorgy Simon, Hui Xiong, Jieping Ye, and Pusheng Zhang. We would also like to thank the students of our data mining classes at the University of Minnesota and Michigan State University who worked with early drafts of the book and provided invaluable feedback. We specifically note the helpful suggestions of Bernardo Craemer, Arifin Ruslim, Jamshid Vayghan, and Yu Wei.

Joydeep Ghosh (University of Texas) and Sanjay Ranka (University of Florida) class tested early versions of the book. We also received many useful





Preface to the Second Edition ix

suggestions directly from the following UT students: Pankaj Adhikari, Rajiv Bhatia, Frederic Bosche, Arindam Chakraborty, Meghana Deodhar, Chris Everson, David Gardner, Saad Godil, Todd Hay, Clint Jones, Ajay Joshi, Joonsoo Lee, Yue Luo, Anuj Nanavati, Tyler Olsen, Sunyoung Park, Aashish Phansalkar, Geoff Prewett, Michael Ryoo, Daryl Shannon, and Mei Yang.

Ronald Kostoff (ONR) read an early version of the clustering chapter and offered numerous suggestions. George Karypis provided invaluable \LaTeX assistance in creating an author index. Irene Moulitsas also provided assistance with \LaTeX and reviewed some of the appendices. Musetta Steinbach was very helpful in finding errors in the figures.

We would like to acknowledge our colleagues at the University of Minnesota and Michigan State who have helped create a positive environment for data mining research. They include Arindam Banerjee, Dan Boley, Joyce Chai, Anil Jain, Ravi Janardan, Rong Jin, George Karypis, Claudia Neuhauser, Haesun Park, William F. Punch, György Simon, Shashi Shekhar, and Jaideep Srivastava. The collaborators on our many data mining projects, who also have our gratitude, include Ramesh Agrawal, Maneesh Bhargava, Steve Cannon, Alok Choudhary, Imme Ebert-Uphoff, Auroop Ganguly, Piet C. de Groen, Fran Hill, Yongdae Kim, Steve Klooster, Kerry Long, Nihar Mahapatra, Rama Nemani, Nikunj Oza, Chris Potter, Lisiane Pruinelli, Nagiza Samatova, Jonathan Shapiro, Kevin Silverstein, Brian Van Ness, Bonnie Westra, Nevin Young, and Zhi-Li Zhang.

The departments of Computer Science and Engineering at the University of Minnesota and Michigan State University provided computing resources and a supportive environment for this project. ARDA, ARL, ARO, DOE, NASA, NOAA, and NSF provided research support for Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. In particular, Kamal Abdali, Mitra Basu, Dick Brackney, Jagdish Chandra, Joe Coughlan, Michael Coyle, Stephen Davis, Frederica Darema, Richard Hirsch, Chandrika Kamath, Tsengdar Lee, Raju Namburu, N. Radhakrishnan, James Sidoran, Sylvia Spengler, Bhavani Thuraisingham, Walt Tiernin, Maria Zemankova, Aidong Zhang, and Xiaodong Zhang have been supportive of our research in data mining and high-performance computing.

It was a pleasure working with the helpful staff at Pearson Education. In particular, we would like to thank Matt Goldstein, Kathy Smith, Carole Snyder, and Joyce Wells. We would also like to thank George Nichols, who helped with the art work and Paul Anagnostopoulos, who provided \LaTeX support.

We are grateful to the following Pearson reviewers: Leman Akoglu (Carnegie Mellon University), Chien-Chung Chan (University of Akron), Zhengxin Chen





x Preface to the Second Edition

(University of Nebraska at Omaha), Chris Clifton (Purdue University), Joydeep Ghosh (University of Texas, Austin), Nazli Goharian (Illinois Institute of Technology), J. Michael Hardin (University of Alabama), Jingrui He (Arizona State University), James Hearne (Western Washington University), Hillol Kargupta (University of Maryland, Baltimore County and Agnik, LLC), Eamonn Keogh (University of California-Riverside), Bing Liu (University of Illinois at Chicago), Mariofanna Milanova (University of Arkansas at Little Rock), Srinivasan Parthasarathy (Ohio State University), Zbigniew W. Ras (University of North Carolina at Charlotte), Xintao Wu (University of North Carolina at Charlotte), and Mohammed J. Zaki (Rensselaer Polytechnic Institute).

Over the years since the first edition, we have also received numerous comments from readers and students who have pointed out typos and various other issues. We are unable to mention these individuals by name, but their input is much appreciated and has been taken into account for the second edition.





Contents

Preface to the Second Edition	v
1 Introduction	1
1.1 What Is Data Mining?	4
1.2 Motivating Challenges	5
1.3 The Origins of Data Mining	7
1.4 Data Mining Tasks	9
1.5 Scope and Organization of the Book	13
1.6 Bibliographic Notes	15
1.7 Exercises	21
2 Data	23
2.1 Types of Data	26
2.1.1 Attributes and Measurement	27
2.1.2 Types of Data Sets	34
2.2 Data Quality	42
2.2.1 Measurement and Data Collection Issues	42
2.2.2 Issues Related to Applications	49
2.3 Data Preprocessing	50
2.3.1 Aggregation	51
2.3.2 Sampling	52
2.3.3 Dimensionality Reduction	56
2.3.4 Feature Subset Selection	58
2.3.5 Feature Creation	61
2.3.6 Discretization and Binarization	63
2.3.7 Variable Transformation	69
2.4 Measures of Similarity and Dissimilarity	71
2.4.1 Basics	72
2.4.2 Similarity and Dissimilarity between Simple Attributes	74
2.4.3 Dissimilarities between Data Objects	76
2.4.4 Similarities between Data Objects	78





xii Contents

2.4.5	Examples of Proximity Measures	79
2.4.6	Mutual Information	88
2.4.7	Kernel Functions*	90
2.4.8	Bregman Divergence*	94
2.4.9	Issues in Proximity Calculation	96
2.4.10	Selecting the Right Proximity Measure	98
2.5	Bibliographic Notes	100
2.6	Exercises	105
3	Classification: Basic Concepts and Techniques	113
3.1	Basic Concepts	114
3.2	General Framework for Classification	117
3.3	Decision Tree Classifier	119
3.3.1	A Basic Algorithm to Build a Decision Tree	121
3.3.2	Methods for Expressing Attribute Test Conditions	124
3.3.3	Measures for Selecting an Attribute Test Condition	127
3.3.4	Algorithm for Decision Tree Induction	136
3.3.5	Example Application: Web Robot Detection	138
3.3.6	Characteristics of Decision Tree Classifiers	140
3.4	Model Overfitting	147
3.4.1	Reasons for Model Overfitting	149
3.5	Model Selection	156
3.5.1	Using a Validation Set	156
3.5.2	Incorporating Model Complexity	157
3.5.3	Estimating Statistical Bounds	162
3.5.4	Model Selection for Decision Trees	162
3.6	Model Evaluation	164
3.6.1	Holdout Method	165
3.6.2	Cross-Validation	165
3.7	Presence of Hyper-parameters	168
3.7.1	Hyper-parameter Selection	168
3.7.2	Nested Cross-Validation	170
3.8	Pitfalls of Model Selection and Evaluation	172
3.8.1	Overlap between Training and Test Sets	172
3.8.2	Use of Validation Error as Generalization Error	172
3.9	Model Comparison*	173
3.9.1	Estimating the Confidence Interval for Accuracy	174
3.9.2	Comparing the Performance of Two Models	175
3.10	Bibliographic Notes	176
3.11	Exercises	185





4	Classification: Alternative Techniques	193
4.1	Types of Classifiers	193
4.2	Rule-Based Classifier	195
4.2.1	How a Rule-Based Classifier Works	197
4.2.2	Properties of a Rule Set	198
4.2.3	Direct Methods for Rule Extraction	199
4.2.4	Indirect Methods for Rule Extraction	204
4.2.5	Characteristics of Rule-Based Classifiers	206
4.3	Nearest Neighbor Classifiers	208
4.3.1	Algorithm	209
4.3.2	Characteristics of Nearest Neighbor Classifiers	210
4.4	Naïve Bayes Classifier	212
4.4.1	Basics of Probability Theory	213
4.4.2	Naïve Bayes Assumption	218
4.5	Bayesian Networks	227
4.5.1	Graphical Representation	227
4.5.2	Inference and Learning	233
4.5.3	Characteristics of Bayesian Networks	242
4.6	Logistic Regression	243
4.6.1	Logistic Regression as a Generalized Linear Model	244
4.6.2	Learning Model Parameters	245
4.6.3	Characteristics of Logistic Regression	248
4.7	Artificial Neural Network (ANN)	249
4.7.1	Perceptron	250
4.7.2	Multi-layer Neural Network	254
4.7.3	Characteristics of ANN	261
4.8	Deep Learning	262
4.8.1	Using Synergistic Loss Functions	263
4.8.2	Using Responsive Activation Functions	266
4.8.3	Regularization	268
4.8.4	Initialization of Model Parameters	271
4.8.5	Characteristics of Deep Learning	275
4.9	Support Vector Machine (SVM)	276
4.9.1	Margin of a Separating Hyperplane	276
4.9.2	Linear SVM	278
4.9.3	Soft-margin SVM	284
4.9.4	Nonlinear SVM	290
4.9.5	Characteristics of SVM	294
4.10	Ensemble Methods	296
4.10.1	Rationale for Ensemble Method	297





xiv Contents

4.10.2	Methods for Constructing an Ensemble Classifier	297
4.10.3	Bias-Variance Decomposition	300
4.10.4	Bagging	302
4.10.5	Boosting	305
4.10.6	Random Forests	310
4.10.7	Empirical Comparison among Ensemble Methods	312
4.11	Class Imbalance Problem	313
4.11.1	Building Classifiers with Class Imbalance	314
4.11.2	Evaluating Performance with Class Imbalance	318
4.11.3	Finding an Optimal Score Threshold	322
4.11.4	Aggregate Evaluation of Performance	323
4.12	Multiclass Problem	330
4.13	Bibliographic Notes	333
4.14	Exercises	345
5	Association Analysis: Basic Concepts and Algorithms	357
5.1	Preliminaries	358
5.2	Frequent Itemset Generation	362
5.2.1	The <i>Apriori</i> Principle	363
5.2.2	Frequent Itemset Generation in the <i>Apriori</i> Algorithm .	364
5.2.3	Candidate Generation and Pruning	368
5.2.4	Support Counting	373
5.2.5	Computational Complexity	377
5.3	Rule Generation	380
5.3.1	Confidence-Based Pruning	380
5.3.2	Rule Generation in <i>Apriori</i> Algorithm	381
5.3.3	An Example: Congressional Voting Records	382
5.4	Compact Representation of Frequent Itemsets	384
5.4.1	Maximal Frequent Itemsets	384
5.4.2	Closed Itemsets	386
5.5	Alternative Methods for Generating Frequent Itemsets*	389
5.6	FP-Growth Algorithm*	393
5.6.1	FP-Tree Representation	394
5.6.2	Frequent Itemset Generation in FP-Growth Algorithm .	397
5.7	Evaluation of Association Patterns	401
5.7.1	Objective Measures of Interestingness	402
5.7.2	Measures beyond Pairs of Binary Variables	414
5.7.3	Simpson's Paradox	416
5.8	Effect of Skewed Support Distribution	418
5.9	Bibliographic Notes	424





5.10 Exercises	438
6 Association Analysis: Advanced Concepts	451
6.1 Handling Categorical Attributes	451
6.2 Handling Continuous Attributes	454
6.2.1 Discretization-Based Methods	454
6.2.2 Statistics-Based Methods	458
6.2.3 Non-discretization Methods	460
6.3 Handling a Concept Hierarchy	462
6.4 Sequential Patterns	464
6.4.1 Preliminaries	465
6.4.2 Sequential Pattern Discovery	468
6.4.3 Timing Constraints*	473
6.4.4 Alternative Counting Schemes*	477
6.5 Subgraph Patterns	479
6.5.1 Preliminaries	480
6.5.2 Frequent Subgraph Mining	483
6.5.3 Candidate Generation	487
6.5.4 Candidate Pruning	493
6.5.5 Support Counting	493
6.6 Infrequent Patterns*	493
6.6.1 Negative Patterns	494
6.6.2 Negatively Correlated Patterns	495
6.6.3 Comparisons among Infrequent Patterns, Negative Pat- terns, and Negatively Correlated Patterns	496
6.6.4 Techniques for Mining Interesting Infrequent Patterns .	498
6.6.5 Techniques Based on Mining Negative Patterns	499
6.6.6 Techniques Based on Support Expectation	501
6.7 Bibliographic Notes	505
6.8 Exercises	510
7 Cluster Analysis: Basic Concepts and Algorithms	525
7.1 Overview	528
7.1.1 What Is Cluster Analysis?	528
7.1.2 Different Types of Clusterings	529
7.1.3 Different Types of Clusters	531
7.2 K-means	534
7.2.1 The Basic K-means Algorithm	535
7.2.2 K-means: Additional Issues	544
7.2.3 Bisecting K-means	547





xvi Contents

7.2.4	K-means and Different Types of Clusters	548
7.2.5	Strengths and Weaknesses	549
7.2.6	K-means as an Optimization Problem	549
7.3	Agglomerative Hierarchical Clustering	554
7.3.1	Basic Agglomerative Hierarchical Clustering Algorithm	555
7.3.2	Specific Techniques	557
7.3.3	The Lance-Williams Formula for Cluster Proximity . . .	562
7.3.4	Key Issues in Hierarchical Clustering	563
7.3.5	Outliers	564
7.3.6	Strengths and Weaknesses	565
7.4	DBSCAN	565
7.4.1	Traditional Density: Center-Based Approach	565
7.4.2	The DBSCAN Algorithm	567
7.4.3	Strengths and Weaknesses	569
7.5	Cluster Evaluation	571
7.5.1	Overview	571
7.5.2	Unsupervised Cluster Evaluation Using Cohesion and Separation	574
7.5.3	Unsupervised Cluster Evaluation Using the Proximity Matrix	582
7.5.4	Unsupervised Evaluation of Hierarchical Clustering . . .	585
7.5.5	Determining the Correct Number of Clusters	587
7.5.6	Clustering Tendency	588
7.5.7	Supervised Measures of Cluster Validity	589
7.5.8	Assessing the Significance of Cluster Validity Measures .	594
7.5.9	Choosing a Cluster Validity Measure	596
7.6	Bibliographic Notes	597
7.7	Exercises	603
8	Cluster Analysis: Additional Issues and Algorithms	613
8.1	Characteristics of Data, Clusters, and Clustering Algorithms .	614
8.1.1	Example: Comparing K-means and DBSCAN	614
8.1.2	Data Characteristics	615
8.1.3	Cluster Characteristics	617
8.1.4	General Characteristics of Clustering Algorithms	619
8.2	Prototype-Based Clustering	621
8.2.1	Fuzzy Clustering	621
8.2.2	Clustering Using Mixture Models	627
8.2.3	Self-Organizing Maps (SOM)	637
8.3	Density-Based Clustering	644





8.3.1	Grid-Based Clustering	644
8.3.2	Subspace Clustering	648
8.3.3	DENCLUE: A Kernel-Based Scheme for Density-Based Clustering	652
8.4	Graph-Based Clustering	656
8.4.1	Sparsification	657
8.4.2	Minimum Spanning Tree (MST) Clustering	658
8.4.3	OPOSSUM: Optimal Partitioning of Sparse Similarities Using METIS	659
8.4.4	Chameleon: Hierarchical Clustering with Dynamic Modeling	660
8.4.5	Spectral Clustering	666
8.4.6	Shared Nearest Neighbor Similarity	673
8.4.7	The Jarvis-Patrick Clustering Algorithm	676
8.4.8	SNN Density	678
8.4.9	SNN Density-Based Clustering	679
8.5	Scalable Clustering Algorithms	681
8.5.1	Scalability: General Issues and Approaches	681
8.5.2	BIRCH	684
8.5.3	CURE	686
8.6	Which Clustering Algorithm?	690
8.7	Bibliographic Notes	693
8.8	Exercises	699
9	Anomaly Detection	703
9.1	Characteristics of Anomaly Detection Problems	705
9.1.1	A Definition of an Anomaly	705
9.1.2	Nature of Data	706
9.1.3	How Anomaly Detection is Used	707
9.2	Characteristics of Anomaly Detection Methods	708
9.3	Statistical Approaches	710
9.3.1	Using Parametric Models	710
9.3.2	Using Non-parametric Models	714
9.3.3	Modeling Normal and Anomalous Classes	715
9.3.4	Assessing Statistical Significance	717
9.3.5	Strengths and Weaknesses	718
9.4	Proximity-based Approaches	719
9.4.1	Distance-based Anomaly Score	719
9.4.2	Density-based Anomaly Score	720
9.4.3	Relative Density-based Anomaly Score	722





xviii Contents

9.4.4	Strengths and Weaknesses	723
9.5	Clustering-based Approaches	724
9.5.1	Finding Anomalous Clusters	724
9.5.2	Finding Anomalous Instances	725
9.5.3	Strengths and Weaknesses	728
9.6	Reconstruction-based Approaches	728
9.6.1	Strengths and Weaknesses	731
9.7	One-class Classification	732
9.7.1	Use of Kernels	733
9.7.2	The Origin Trick	734
9.7.3	Strengths and Weaknesses	738
9.8	Information Theoretic Approaches	738
9.8.1	Strengths and Weaknesses	740
9.9	Evaluation of Anomaly Detection	740
9.10	Bibliographic Notes	742
9.11	Exercises	749
10	Avoiding False Discoveries	755
10.1	Preliminaries: Statistical Testing	756
10.1.1	Significance Testing	756
10.1.2	Hypothesis Testing	761
10.1.3	Multiple Hypothesis Testing	767
10.1.4	Pitfalls in Statistical Testing	776
10.2	Modeling Null and Alternative Distributions	778
10.2.1	Generating Synthetic Data Sets	781
10.2.2	Randomizing Class Labels	782
10.2.3	Resampling Instances	782
10.2.4	Modeling the Distribution of the Test Statistic	783
10.3	Statistical Testing for Classification	783
10.3.1	Evaluating Classification Performance	783
10.3.2	Binary Classification as Multiple Hypothesis Testing	785
10.3.3	Multiple Hypothesis Testing in Model Selection	786
10.4	Statistical Testing for Association Analysis	787
10.4.1	Using Statistical Models	788
10.4.2	Using Randomization Methods	794
10.5	Statistical Testing for Cluster Analysis	795
10.5.1	Generating a Null Distribution for Internal Indices	796
10.5.2	Generating a Null Distribution for External Indices	798
10.5.3	Enrichment	798
10.6	Statistical Testing for Anomaly Detection	800





Contents xix

10.7 Bibliographic Notes	803
10.8 Exercises	808
Author Index	816
Subject Index	829
Copyright Permissions	839



