

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

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

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

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

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 Patterns, 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

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

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

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