

# Data Mining: Concepts and Techniques

Jiawei Han

Micheline Kamber

*Simon Fraser University*

Technische Universität Darmstadt	
FACHBEREICH INFORMATIK	
<b>B I B L I O T H E K</b>	
Inventar-Nr.:	<u>101-00461</u>
Sachgebiete:	_____
Standort:	_____



**MORGAN KAUFMANN PUBLISHERS**

AN IMPRINT OF ACADEMIC PRESS

A Harcourt Science and Technology Company

SAN FRANCISCO SAN DIEGO NEW YORK BOSTON  
LONDON SYDNEY TOKYO

# Contents

**Foreword vii**

**Preface xix**

## Chapter 1 **Introduction 1**

- 1.1 **What Motivated Data Mining? Why Is It Important? 1**
- 1.2 **So, What Is Data Mining? 5**
- 1.3 **Data Mining—On What Kind of Data? 10**
  - 1.3.1 Relational Databases 10
  - 1.3.2 Data Warehouses 12
  - 1.3.3 Transactional Databases 15
  - 1.3.4 Advanced Database Systems and Advanced Database Applications 16
- 1.4 **Data Mining Functionalities—What Kinds of Patterns Can Be Mined? 21**
  - 1.4.1 Concept/Class Description: Characterization and Discrimination 21
  - 1.4.2 Association Analysis 23
  - 1.4.3 Classification and Prediction 24
  - 1.4.4 Cluster Analysis 25
  - 1.4.5 Outlier Analysis 25
  - 1.4.6 Evolution Analysis 26
- 1.5 **Are All of the Patterns Interesting? 27**
- 1.6 **Classification of Data Mining Systems 28**
- 1.7 **Major Issues in Data Mining 30**
- 1.8 **Summary 33**
- Exercises 34**
- Bibliographic Notes 35**

## Chapter 2 **Data Warehouse and OLAP Technology for Data Mining 39**

- 2.1 **What Is a Data Warehouse? 39**
  - 2.1.1 Differences between Operational Database Systems and Data Warehouses 42
  - 2.1.2 But, Why Have a Separate Data Warehouse? 44

2.2	<b>A Multidimensional Data Model</b>	<b>44</b>
2.2.1	From Tables and Spreadsheets to Data Cubes	45
2.2.2	Stars, Snowflakes, and Fact Constellations: Schemas for Multidimensional Databases	48
2.2.3	Examples for Defining Star, Snowflake, and Fact Constellation Schemas	52
2.2.4	Measures: Their Categorization and Computation	54
2.2.5	Introducing Concept Hierarchies	56
2.2.6	OLAP Operations in the Multidimensional Data Model	58
2.2.7	A Starnet Query Model for Querying Multidimensional Databases	61
2.3	<b>Data Warehouse Architecture</b>	<b>62</b>
2.3.1	Steps for the Design and Construction of Data Warehouses	63
2.3.2	A Three-Tier Data Warehouse Architecture	65
2.3.3	Types of OLAP Servers: ROLAP versus MOLAP versus HOLAP	69
2.4	<b>Data Warehouse Implementation</b>	<b>71</b>
2.4.1	Efficient Computation of Data Cubes	71
2.4.2	Indexing OLAP Data	79
2.4.3	Efficient Processing of OLAP Queries	81
2.4.4	Metadata Repository	83
2.4.5	Data Warehouse Back-End Tools and Utilities	84
2.5	<b>Further Development of Data Cube Technology</b>	<b>85</b>
2.5.1	Discovery-Driven Exploration of Data Cubes	85
2.5.2	Complex Aggregation at Multiple Granularities: Multifeature Cubes	89
2.5.3	Other Developments	92
2.6	<b>From Data Warehousing to Data Mining</b>	<b>93</b>
2.6.1	Data Warehouse Usage	93
2.6.2	From On-Line Analytical Processing to On-Line Analytical Mining	95
2.7	<b>Summary</b>	<b>98</b>
	<b>Exercises</b>	<b>99</b>
	<b>Bibliographic Notes</b>	<b>103</b>
Chapter 3	<b>Data Preprocessing</b>	<b>105</b>
3.1	<b>Why Preprocess the Data?</b>	<b>105</b>
3.2	<b>Data Cleaning</b>	<b>109</b>
3.2.1	Missing Values	109
3.2.2	Noisy Data	110
3.2.3	Inconsistent Data	112
3.3	<b>Data Integration and Transformation</b>	<b>112</b>
3.3.1	Data Integration	112
3.3.2	Data Transformation	114

3.4	<b>Data Reduction</b>	<b>116</b>
3.4.1	Data Cube Aggregation	117
3.4.2	Dimensionality Reduction	119
3.4.3	Data Compression	121
3.4.4	Numerosity Reduction	124
3.5	<b>Discretization and Concept Hierarchy Generation</b>	<b>130</b>
3.5.1	Discretization and Concept Hierarchy Generation for Numeric Data	132
3.5.2	Concept Hierarchy Generation for Categorical Data	138
3.6	<b>Summary</b>	<b>140</b>
	<b>Exercises</b>	<b>141</b>
	<b>Bibliographic Notes</b>	<b>142</b>
Chapter 4	<b>Data Mining Primitives, Languages, and System Architectures</b>	<b>145</b>
4.1	<b>Data Mining Primitives: What Defines a Data Mining Task?</b>	<b>146</b>
4.1.1	Task-Relevant Data	148
4.1.2	The Kind of Knowledge to be Mined	150
4.1.3	Background Knowledge: Concept Hierarchies	151
4.1.4	Interestingness Measures	155
4.1.5	Presentation and Visualization of Discovered Patterns	157
4.2	<b>A Data Mining Query Language</b>	<b>159</b>
4.2.1	Syntax for Task-Relevant Data Specification	160
4.2.2	Syntax for Specifying the Kind of Knowledge to be Mined	162
4.2.3	Syntax for Concept Hierarchy Specification	165
4.2.4	Syntax for Interestingness Measure Specification	166
4.2.5	Syntax for Pattern Presentation and Visualization Specification	167
4.2.6	Putting It All Together—An Example of a DMQL Query	167
4.2.7	Other Data Mining Languages and the Standardization of Data Mining Primitives	169
4.3	<b>Designing Graphical User Interfaces Based on a Data Mining Query Language</b>	<b>170</b>
4.4	<b>Architectures of Data Mining Systems</b>	<b>171</b>
4.5	<b>Summary</b>	<b>174</b>
	<b>Exercises</b>	<b>174</b>
	<b>Bibliographic Notes</b>	<b>176</b>
Chapter 5	<b>Concept Description: Characterization and Comparison</b>	<b>179</b>
5.1	<b>What Is Concept Description?</b>	<b>179</b>
5.2	<b>Data Generalization and Summarization-Based Characterization</b>	<b>181</b>

5.2.1	Attribute-Oriented Induction	182
5.2.2	Efficient Implementation of Attribute-Oriented Induction	187
5.2.3	Presentation of the Derived Generalization	190
5.3	<b>Analytical Characterization: Analysis of Attribute Relevance</b>	<b>194</b>
5.3.1	Why Perform Attribute Relevance Analysis?	195
5.3.2	Methods of Attribute Relevance Analysis	196
5.3.3	Analytical Characterization: An Example	198
5.4	<b>Mining Class Comparisons: Discriminating between Different Classes</b>	<b>200</b>
5.4.1	Class Comparison Methods and Implementations	201
5.4.2	Presentation of Class Comparison Descriptions	204
5.4.3	Class Description: Presentation of Both Characterization and Comparison	206
5.5	<b>Mining Descriptive Statistical Measures in Large Databases</b>	<b>208</b>
5.5.1	Measuring the Central Tendency	209
5.5.2	Measuring the Dispersion of Data	210
5.5.3	Graph Displays of Basic Statistical Class Descriptions	213
5.6	<b>Discussion</b>	<b>217</b>
5.6.1	Concept Description: A Comparison with Typical Machine Learning Methods	218
5.6.2	Incremental and Parallel Mining of Concept Description	220
5.7	<b>Summary</b>	<b>220</b>
	<b>Exercises</b>	<b>222</b>
	<b>Bibliographic Notes</b>	<b>223</b>
Chapter 6	<b>Mining Association Rules in Large Databases</b>	<b>225</b>
6.1	<b>Association Rule Mining</b>	<b>226</b>
6.1.1	Market Basket Analysis: A Motivating Example for Association Rule Mining	226
6.1.2	Basic Concepts	227
6.1.3	Association Rule Mining: A Road Map	229
6.2	<b>Mining Single-Dimensional Boolean Association Rules from Transactional Databases</b>	<b>230</b>
6.2.1	The Apriori Algorithm: Finding Frequent Itemsets Using Candidate Generation	230
6.2.2	Generating Association Rules from Frequent Itemsets	236
6.2.3	Improving the Efficiency of Apriori	236
6.2.4	Mining Frequent Itemsets without Candidate Generation	239
6.2.5	Iceberg Queries	243
6.3	<b>Mining Multilevel Association Rules from Transaction Databases</b>	<b>244</b>

6.3.1	Multilevel Association Rules	244
6.3.2	Approaches to Mining Multilevel Association Rules	246
6.3.3	Checking for Redundant Multilevel Association Rules	250
6.4	<b>Mining Multidimensional Association Rules from Relational Databases and Data Warehouses</b>	<b>251</b>
6.4.1	Multidimensional Association Rules	251
6.4.2	Mining Multidimensional Association Rules Using Static Discretization of Quantitative Attributes	253
6.4.3	Mining Quantitative Association Rules	254
6.4.4	Mining Distance-Based Association Rules	257
6.5	<b>From Association Mining to Correlation Analysis</b>	<b>259</b>
6.5.1	Strong Rules Are Not Necessarily Interesting: An Example	259
6.5.2	From Association Analysis to Correlation Analysis	260
6.6	<b>Constraint-Based Association Mining</b>	<b>262</b>
6.6.1	Metarule-Guided Mining of Association Rules	263
6.6.2	Mining Guided by Additional Rule Constraints	265
6.7	<b>Summary</b>	<b>269</b>
	<b>Exercises</b>	<b>271</b>
	<b>Bibliographic Notes</b>	<b>276</b>
Chapter 7	<b>Classification and Prediction</b>	<b>279</b>
7.1	<b>What Is Classification? What Is Prediction?</b>	<b>279</b>
7.2	<b>Issues Regarding Classification and Prediction</b>	<b>282</b>
7.2.1	Preparing the Data for Classification and Prediction	282
7.2.2	Comparing Classification Methods	283
7.3	<b>Classification by Decision Tree Induction</b>	<b>284</b>
7.3.1	Decision Tree Induction	285
7.3.2	Tree Pruning	289
7.3.3	Extracting Classification Rules from Decision Trees	290
7.3.4	Enhancements to Basic Decision Tree Induction	291
7.3.5	Scalability and Decision Tree Induction	292
7.3.6	Integrating Data Warehousing Techniques and Decision Tree Induction	294
7.4	<b>Bayesian Classification</b>	<b>296</b>
7.4.1	Bayes Theorem	296
7.4.2	Naive Bayesian Classification	297
7.4.3	Bayesian Belief Networks	299
7.4.4	Training Bayesian Belief Networks	301
7.5	<b>Classification by Backpropagation</b>	<b>303</b>
7.5.1	A Multilayer Feed-Forward Neural Network	303
7.5.2	Defining a Network Topology	304

7.5.3	Backpropagation	305
7.5.4	Backpropagation and Interpretability	310
7.6	<b>Classification Based on Concepts from Association Rule Mining</b>	<b>311</b>
7.7	<b>Other Classification Methods</b>	<b>314</b>
7.7.1	k-Nearest Neighbor Classifiers	314
7.7.2	Case-Based Reasoning	315
7.7.3	Genetic Algorithms	316
7.7.4	Rough Set Approach	316
7.7.5	Fuzzy Set Approaches	317
7.8	<b>Prediction</b>	<b>319</b>
7.8.1	Linear and Multiple Regression	319
7.8.2	Nonlinear Regression	321
7.8.3	Other Regression Models	322
7.9	<b>Classifier Accuracy</b>	<b>322</b>
7.9.1	Estimating Classifier Accuracy	323
7.9.2	Increasing Classifier Accuracy	324
7.9.3	Is Accuracy Enough to Judge a Classifier?	325
7.10	<b>Summary</b>	<b>326</b>
	<b>Exercises</b>	<b>328</b>
	<b>Bibliographic Notes</b>	<b>330</b>

## Chapter 8 **Cluster Analysis** 335

8.1	<b>What Is Cluster Analysis?</b>	<b>335</b>
8.2	<b>Types of Data in Cluster Analysis</b>	<b>338</b>
8.2.1	Interval-Scaled Variables	339
8.2.2	Binary Variables	341
8.2.3	Nominal, Ordinal, and Ratio-Scaled Variables	343
8.2.4	Variables of Mixed Types	345
8.3	<b>A Categorization of Major Clustering Methods</b>	<b>346</b>
8.4	<b>Partitioning Methods</b>	<b>348</b>
8.4.1	Classical Partitioning Methods: k-Means and k-Medoids	349
8.4.2	Partitioning Methods in Large Databases: From k-Medoids to CLARANS	353
8.5	<b>Hierarchical Methods</b>	<b>354</b>
8.5.1	Agglomerative and Divisive Hierarchical Clustering	355
8.5.2	BIRCH: Balanced Iterative Reducing and Clustering Using Hierarchies	357
8.5.3	CURE: Clustering Using REpresentatives	358
8.5.4	Chameleon: A Hierarchical Clustering Algorithm Using Dynamic Modeling	361

8.6	<b>Density-Based Methods</b>	<b>363</b>
8.6.1	DBSCAN: A Density-Based Clustering Method Based on Connected Regions with Sufficiently High Density	363
8.6.2	OPTICS: Ordering Points To Identify the Clustering Structure	365
8.6.3	DENCLUE: Clustering Based on Density Distribution Functions	366
8.7	<b>Grid-Based Methods</b>	<b>370</b>
8.7.1	STING: STatistical INformation Grid	370
8.7.2	WaveCluster: Clustering Using Wavelet Transformation	372
8.7.3	CLIQUE: Clustering High-Dimensional Space	374
8.8	<b>Model-Based Clustering Methods</b>	<b>376</b>
8.8.1	Statistical Approach	376
8.8.2	Neural Network Approach	379
8.9	<b>Outlier Analysis</b>	<b>381</b>
8.9.1	Statistical-Based Outlier Detection	382
8.9.2	Distance-Based Outlier Detection	384
8.9.3	Deviation-Based Outlier Detection	386
8.10	<b>Summary</b>	<b>388</b>
	<b>Exercises</b>	<b>389</b>
	<b>Bibliographic Notes</b>	<b>391</b>

## Chapter 9 **Mining Complex Types of Data** 395

9.1	<b>Multidimensional Analysis and Descriptive Mining of Complex Data Objects</b>	<b>396</b>
9.1.1	Generalization of Structured Data	396
9.1.2	Aggregation and Approximation in Spatial and Multimedia Data Generalization	397
9.1.3	Generalization of Object Identifiers and Class/Subclass Hierarchies	399
9.1.4	Generalization of Class Composition Hierarchies	399
9.1.5	Construction and Mining of Object Cubes	400
9.1.6	Generalization-Based Mining of Plan Databases by Divide-and-Conquer	401
9.2	<b>Mining Spatial Databases</b>	<b>405</b>
9.2.1	Spatial Data Cube Construction and Spatial OLAP	405
9.2.2	Spatial Association Analysis	410
9.2.3	Spatial Clustering Methods	411
9.2.4	Spatial Classification and Spatial Trend Analysis	411
9.2.5	Mining Raster Databases	412
9.3	<b>Mining Multimedia Databases</b>	<b>412</b>
9.3.1	Similarity Search in Multimedia Data	412
9.3.2	Multidimensional Analysis of Multimedia Data	414
9.3.3	Classification and Prediction Analysis of Multimedia Data	416



	9.3.4 Mining Associations in Multimedia Data	417
9.4	<b>Mining Time-Series and Sequence Data</b>	<b>418</b>
	9.4.1 Trend Analysis	418
	9.4.2 Similarity Search in Time-Series Analysis	421
	9.4.3 Sequential Pattern Mining	424
	9.4.4 Periodicity Analysis	426
9.5	<b>Mining Text Databases</b>	<b>428</b>
	9.5.1 Text Data Analysis and Information Retrieval	428
	9.5.2 Text Mining: Keyword-Based Association and Document Classification	433
9.6	<b>Mining the World Wide Web</b>	<b>435</b>
	9.6.1 Mining the Web's Link Structures to Identify Authoritative Web Pages	437
	9.6.2 Automatic Classification of Web Documents	439
	9.6.3 Construction of a Multilayered Web Information Base	440
	9.6.4 Web Usage Mining	441
9.7	<b>Summary</b>	<b>443</b>
	<b>Exercises</b>	<b>444</b>
	<b>Bibliographic Notes</b>	<b>446</b>
Chapter 10	<b>Applications and Trends in Data Mining</b>	<b>451</b>
10.1	<b>Data Mining Applications</b>	<b>451</b>
	10.1.1 Data Mining for Biomedical and DNA Data Analysis	451
	10.1.2 Data Mining for Financial Data Analysis	453
	10.1.3 Data Mining for the Retail Industry	455
	10.1.4 Data Mining for the Telecommunication Industry	456
10.2	<b>Data Mining System Products and Research Prototypes</b>	<b>457</b>
	10.2.1 How to Choose a Data Mining System	458
	10.2.2 Examples of Commercial Data Mining Systems	461
10.3	<b>Additional Themes on Data Mining</b>	<b>462</b>
	10.3.1 Visual and Audio Data Mining	462
	10.3.2 Scientific and Statistical Data Mining	464
	10.3.3 Theoretical Foundations of Data Mining	470
	10.3.4 Data Mining and Intelligent Query Answering	471
10.4	<b>Social Impacts of Data Mining</b>	<b>472</b>
	10.4.1 Is Data Mining a Hype or a Persistent, Steadily Growing Business?	473
	10.4.2 Is Data Mining Merely Managers' Business or Everyone's Business?	475
	10.4.3 Is Data Mining a Threat to Privacy and Data Security?	476
10.5	<b>Trends in Data Mining</b>	<b>478</b>

10.6	<b>Summary</b>	<b>480</b>
	<b>Exercises</b>	<b>481</b>
	<b>Bibliographic Notes</b>	<b>483</b>

Appendix A **An Introduction to Microsoft's OLE DB for Data Mining 485**

A.1	<b>Creating a DMM object</b>	<b>486</b>
A.2	<b>Inserting Training Data into the Model and Training the Model</b>	<b>488</b>
A.3	<b>Using the Model</b>	<b>488</b>

Appendix B **An Introduction to DBMiner 493**

B.1	<b>System Architecture</b>	<b>494</b>
B.2	<b>Input and Output</b>	<b>494</b>
B.3	<b>Data Mining Tasks Supported by the System</b>	<b>495</b>
B.4	<b>Support for Task and Method Selection</b>	<b>498</b>
B.5	<b>Support of the KDD Process</b>	<b>499</b>
B.6	<b>Main Applications</b>	<b>499</b>
B.7	<b>Current Status</b>	<b>499</b>

**Bibliography 501**

**Index 533**