

Java Data Mining: Strategy, Standard, and Practice

The Morgan Kaufmann Series in Data Management Systems

Series Editor: Jim Gray, Microsoft Research

Java Data Mining: Strategy, Standard, and Practice
Mark Hornick, Erik Marcade and Sunil Venkayala

Joe Celko's Analytics and OLAP in SQL
Joe Celko

Data Preparation for Data Mining Using SAS
Mamdouh Refaat

Querying XML: XQuery, XPath, and SQL/XML in Context
Jim Melton and Stephen Buxton

Data Mining: Concepts and Techniques, Second Edition
Jiawei Han and Micheline Kamber

Database Modeling and Design: Logical Design, Fourth Edition
Toby J. Teorey, Sam S. Lightstone, and Thomas P. Nadeau

Foundations of Multidimensional and Metric Data Structures
Hanan Samet

Joe Celko's SQL for Smarties: Advanced SQL Programming, Third Edition
Joe Celko

Moving Objects Databases
Ralf Hartmut Güting and Markus Schneider

Joe Celko's SQL Programming Style
Joe Celko

Data Mining, Second Edition: Concepts and Techniques
Ian Witten and Eibe Frank

Fuzzy Modeling and Genetic Algorithms for Data Mining and Exploration
Earl Cox

Data Modeling Essentials, Third Edition
Graeme C. Simsimon and Graham C. Witt

Location-Based Services
Jochen Schiller and Agnès Voisard

Database Modeling with Microsoft® Visio for Enterprise Architects
Terry Halpin, Ken Evans, Patrick Hallock, Bill Maclean

Designing Data-Intensive Web Applications
Stephano Ceri, Piero Fraternali, Aldo Bongio, Marco Brambilla, Sara Comai, and Maristella Matera

Mining the Web: Discovering Knowledge from Hypertext Data
Soumen Chakrabarti

Advanced SQL: 1999—Understanding Object-Relational and Other Advanced Features
Jim Melton

Database Tuning: Principles, Experiments, and Troubleshooting Techniques
Dennis Shasha and Philippe Bonnet

SQL: 1999—Understanding Relational Language Components
Jim Melton and Alan R. Simon

Information Visualization in Data Mining and Knowledge Discovery
Edited by Usama Fayyad, Georges G. Grinstein, and Andreas Wierse

Transactional Information Systems: Theory, Algorithms, and Practice of Concurrency Control and Recovery
Gerhard Weikum and Gottfried Vossen

Spatial Databases: With Application to GIS
Philippe Rigaux, Michel Scholl, and Agnès Voisard

Information Modeling and Relational Databases: From Conceptual Analysis to Logical Design
Terry Halpin

Component Database Systems
Edited by Klaus R. Dittrich and Andreas Geppert

Managing Reference Data in Enterprise Databases: Binding Corporate Data to the Wider World
Malcolm Chisholm

Understanding SQL and Java Together: A Guide to SQLJ, JDBC, and Related Technologies
Jim Melton and Andrew Eisenberg

Database: Principles, Programming, and Performance, Second Edition
Patrick and Elizabeth O'Neil

The Object Data Standard: ODMG 3.0
Edited by R. G. G. Cattell and Douglas K. Barry

Data on the Web: From Relations to Semistructured Data and XML
Serge Abiteboul, Peter Buneman, and Dan Suciu

Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations
Ian Witten and Eibe Frank

Joe Celko's SQL for Smarties: Advanced SQL Programming, Second Edition
Joe Celko

Joe Celko's Data and Databases: Concepts in Practice
Joe Celko

Developing Time-Oriented Database Applications in SQL
Richard T. Snodgrass

Web Farming for the Data Warehouse
Richard D. Hackathorn

Management of Heterogeneous and Autonomous Database Systems
Edited by Ahmed Elmagarmid, Marek Rusinkiewicz, and Amit Sheth

Object-Relational DBMSs: Tracking the Next Great Wave, Second Edition
Michael Stonebraker and Paul Brown, with Dorothy Moore

A Complete Guide to DB2 Universal Database
Don Chamberlin

Universal Database Management: A Guide to Object-Relational Technology
Cynthia Maro Saracco

Readings in Database Systems, Third Edition
Edited by Michael Stonebraker and Joseph M. Hellerstein

Understanding SQL's Stored Procedures: A Complete Guide to SQL/PSM
Jim Melton

Principles of Multimedia Database Systems
V. S. Subrahmanian

Principles of Database Query Processing for Advanced Applications
Clement T. Yu and Weiyi Meng

Advanced Database Systems
Carlo Zaniolo, Stefano Ceri, Christos Faloutsos, Richard T. Snodgrass, V. S. Subrahmanian, and Roberto Zicari

Principles of Transaction Processing
Philip A. Bernstein and Eric Newcomer

Using the New DB2: IBM's Object-Relational Database System
Don Chamberlin

Distributed Algorithms
Nancy A. Lynch

Active Database Systems: Triggers and Rules For Advanced Database Processing
Edited by Jennifer Widom and Stefano Ceri

Migrating Legacy Systems: Gateways, Interfaces, & the Incremental Approach
Michael L. Brodie and Michael Stonebraker

Atomic Transactions
Nancy Lynch, Michael Merritt, William Weihl, and Alan Fekete

Query Processing for Advanced Database Systems
Edited by Johann Christoph Freytag, David Maier, and Gottfried Vossen

Transaction Processing: Concepts and Techniques
Jim Gray and Andreas Reuter

Building an Object-Oriented Database System: The Story of O₂
Edited by François Bancilhon, Claude Delobel, and Paris Kanellakis

Database Transaction Models for Advanced Applications
Edited by Ahmed K. Elmagarmid

A Guide to Developing Client/Server SQL Applications
Setrag Khoshafian, Arvola Chan, Anna Wong, and Harry K. T. Wong

The Benchmark Handbook for Database and Transaction Processing Systems, Second Edition
Edited by Jim Gray

Camelot and Avalon: A Distributed Transaction Facility
Edited by Jeffrey L. Eppinger, Lily B. Mummert, and Alfred Z. Spector

Readings in Object-Oriented Database Systems
Edited by Stanley B. Zdonik and David Maier

Java Data Mining: Strategy, Standard, and Practice

A Practical Guide for Architecture,
Design, and Implementation

Mark F. Hornick
Erik Marcadé
Sunil Venkayala



ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO



MORGAN KAUFMANN PUBLISHERS

Publisher Diane D. Cerra
Publishing Services Manager George Morrison
Project Manager Marilyn E. Rash
Assistant Editor Asma Palmeiro
Cover Design Brian May, Maycreate LLC
Production Services Graphic World
Interior Printer The Maple Press Company
Cover Printer Phoenix Color Corp.

Morgan Kaufmann Publishers is an imprint of Elsevier.
500 Sansome Street, Suite 400, San Francisco, CA 94111

This book is printed on acid-free paper.

© 2007 by Elsevier Inc. All rights reserved.

Designations used by companies to distinguish their products are often claimed as trademarks or registered trademarks. In all instances in which Morgan Kaufmann Publishers is aware of a claim, the product names appear in initial capital or all capital letters. Readers, however, should contact the appropriate companies for more complete information regarding trademarks and registration.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopying, scanning, or otherwise—without prior written permission of the publisher.

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone: (+44) 1865 843830, fax: (+44) 1865 853333, e-mail: permissions@elsevier.com. You may also complete your request on-line via the Elsevier homepage (<http://elsevier.com>), by selecting "Support & Contact" then "Copyright and Permission" and then "Obtaining Permissions."

Java Specification Request 73. Copyright © 2004. Oracle Corporation. Used with permission.
Java Specification Request 274. Copyright © 2005. Oracle Corporation. Used with permission.

Library of Congress Cataloging-in-Publication Data
Application Submitted

ISBN 10: 0-12-370452-9
ISBN 13: 978-0-12-370452-8

For information on all Morgan Kaufmann publications,
visit our Web site at www.mkp.com or www.books.elsevier.com

Printed in the United States of America
06 07 08 09 10 10 9 8 7 6 5 4 3 2 1

Working together to grow
libraries in developing countries

www.elsevier.com | www.bookaid.org | www.sabre.org

ELSEVIER BOOK AID
 International Sabre Foundation

To Suzanne, Amanda, and Tim for their enthusiasm and support. - M.H.

To Caroline, Laetitia and Guillaume. - E.M.

To my parents, wife Meera and daughter Shreya. - S.V.

Contents

Preface	xvii
Guide to Readers	xxi
Part I – Strategy	I
Chapter 1 Overview of Data Mining	3
1.1 Why Data Mining Is Relevant Today?	4
1.2 Introducing Data Mining	6
1.2.1 Data Mining by Other Names	6
1.2.2 Data Mining Versus Other Forms of Advanced Analytics	7
1.2.3 Process	10
1.2.4 What Is a Data Mining Model?	12
1.2.5 Some Jargon	13
<i>The Mining Metaphor</i>	15
1.3 The Value of Data Mining	20
1.3.1 How Reliable Is Data Mining?	20
1.3.2 How Can Data Mining Increase Profits and Reduce Costs	21
1.4 Summary	23
References	24
Chapter 2 Solving Problems in Industry	25
2.1 Cross-Industry Data Mining Solutions	26
2.1.1 Customer Acquisition	26
2.1.2 Customer Retention	28
2.1.3 Response Modeling	30
2.1.4 Fraud Detection	32
2.1.5 Cross-Selling	35

- 2.1.6 New Product Line Development 36
- 2.1.7 Survey Analysis 37
- 2.1.8 Credit Scoring 38
- 2.1.9 Warranty Analysis 39
- 2.1.10 Defect Analysis 40
- 2.2 Data Mining in Industries 41
 - 2.2.1 Financial Services 41
 - 2.2.2 Healthcare 42
 - 2.2.3 Higher Education 43
 - 2.2.4 Public Sector 44
 - 2.2.5 Communications 45
 - 2.2.6 Retail 46
 - 2.2.7 Life Sciences 46
- 2.3 Summary 47
- References 47

Chapter 3 Data Mining Process 51

- 3.1 A Standardized Data Mining Process 52
 - 3.1.1 Business Understanding Phase 53
 - 3.1.2 Data Understanding Phase 55
 - 3.1.3 Data Preparation Phase 56
 - 3.1.4 Modeling Phase 57
 - 3.1.5 Evaluation Phase 58
 - 3.1.6 Deployment Phase 59
- 3.2 A More Detailed View of Data Analysis and Preparation 60
- 3.3 Data Mining Modeling, Analysis, and Scoring Processes 70
 - 3.3.1 Model Building 70
 - 3.3.2 Model Apply 71
 - 3.3.3 Model Test 72
- 3.4 The Role of Databases and Data Warehouses in Data Mining 74
- 3.5 Data Mining in Enterprise Software Architectures 75
 - 3.5.1 Architectures 76
 - 3.5.2 Incorporating Data Mining into Business Operations 79
 - 3.5.3 Business Workflow 80
- 3.6 Advances in Automated Data Mining 81
- 3.7 Summary 82
- References 83

Chapter 4 Mining Functions and Algorithms 85

- 4.1 Data Mining Functions 86
- 4.2 Classification 88

4.3 Regression	89
4.4 Attribute Importance	91
4.5 Association	93
4.6 Clustering	97
4.7 Summary	100
References	101
Chapter 5 JDM Strategy	103
5.1 What Is the JDM Strategy?	104
5.2 Role of Standards	110
5.2.1 Why Create a Standard?	110
5.2.2 What Do Data Mining Standards Enable?	112
5.3 Summary	114
References	114
Chapter 6 Getting Started	117
6.1 Business Understanding	118
6.2 Data Understanding	119
6.3 Data Preparation	121
6.4 Modeling	123
6.4.1 Build	124
6.4.2 Test	126
6.5 Evaluation	127
6.6 Deployment	127
6.7 Summary	129
References	129
Part II – Standard	131
Chapter 7 Java Data Mining Concepts	133
7.1 Classification Problem	134
7.1.1 Problem Definition: <i>How to reduce customer attrition?</i>	134
7.1.2 Solution Approach: <i>Predict customers who are likely to attrite</i>	134
7.1.3 Data Specification: <i>CUSTOMERS dataset</i>	135
7.1.4 Specify Settings: <i>Fine-tune the solution to the problem</i>	139
7.1.5 Select Algorithm: <i>Find the best fit algorithm</i>	141
7.1.6 Evaluate Model Quality: <i>Compute classification test metrics</i>	150
7.1.7 Apply Model: <i>Obtain prediction results</i>	155

- 7.2 Regression Problem 157
 - 7.2.1 Problem Definition: *How to reduce processing time of residential real-estate appraisals?* 157
 - 7.2.2 Solution Approach: *Property value prediction using regression* 157
 - 7.2.3 Data Specification: *REAL_ESTATE_APPRAISALS dataset* 157
 - 7.2.4 Select Algorithm: *Find the best fit algorithm* 158
 - 7.2.5 Evaluate Model Quality: *Compute regression test metrics* 159
 - 7.2.6 Apply Model: *Obtain prediction results* 159
- 7.3 Attribute Importance 160
 - 7.3.1 Problem Definition: *How to find important customer attributes?* 160
 - 7.3.2 Solution Approach: *Rank attributes according to predictive value* 160
 - 7.3.3 Data Specification, Fine-Tune Settings, and Algorithm Selection 160
 - 7.3.4 Use Model Details: *Explore attribute importance values* 161
- 7.4 Association Rules Problem 162
 - 7.4.1 Problem Definition: *How to Identify cross-sell products for customers?* 162
 - 7.4.2 Solution Approach: *Discover product associations from customer data* 162
 - 7.4.3 Data Specification: *CUSTOMERS and their product purchase data* 163
 - 7.4.4 Fine-Tune Settings: *Filter rules based on rule quality metrics* 163
 - 7.4.5 Use Model Content: *Explore rules from the model* 165
- 7.5 Clustering Problem 165
 - 7.5.1 Problem Definition: *How to understand customer behavior and needs?* 165
 - 7.5.2 Solution Approach: *Find clusters of similar customers* 166
 - 7.5.3 Data Specification and Settings 166
 - 7.5.4 Use Model Details: *Explore clusters* 168
 - 7.5.5 Apply Clustering Model: *Assign new cases to the clusters* 169
- 7.6 Summary 170
- References 170

Chapter 8 Design of the JDM API

- 8.1 Object Modeling of Data Mining Concepts 174
 - 8.1.1 Data Specification Objects 175

8.1.2	Settings Objects	178
8.1.3	Models	183
8.1.4	Test Metrics	184
8.1.5	Tasks	185
8.2	Modular Packages	187
8.3	Connection Architecture	188
8.4	Object Factories	190
8.5	Uniform Resource Identifiers for Datasets	192
8.6	Enumerated Types	192
8.7	Exceptions	194
8.8	Discovering DME Capabilities	196
8.9	Summary	197
	References	197

Chapter 9 Using the JDM API

199

9.1	Connection Interfaces	200
9.1.1	Using the <i>ConnectionFactory</i> Interface	201
9.1.2	Using the <i>Connection</i> Interface	203
9.1.3	Executing Mining Operations	209
9.1.4	Exploring Mining Capabilities	211
9.1.5	Finding DME and JDM Version Information	212
9.1.6	Object List Methods	213
9.1.7	Model and Data Load Methods	213
9.2	Using JDM Enumerations	213
9.3	Using Data Specification Interfaces	214
9.4	Using Classification Interfaces	218
9.4.1	Classification Settings	218
9.4.2	Algorithm Settings	220
9.4.3	Model Contents	222
9.4.4	Test Metrics for Model Evaluation	227
9.4.5	Applying a Model to Data in Batch	229
9.4.6	Applying a Model to a Single Record – Real-Time Scoring	234
9.5	Using Regression Interfaces	235
9.6	Using Attribute Importance Interfaces	240
9.7	Using Association Interfaces	243
9.8	Using Clustering Interfaces	249
9.9	Summary	256
	References	257

Chapter 10 XML Schema 259

- 10.1 Overview 260
- 10.2 Schema Elements 260
- 10.3 Schema Types 262
- 10.4 Using PMML with the JDM Schema 267
- 10.5 Use Cases for JDM XML Schema and Documents 270
- 10.6 Summary 271
- References 271

Chapter 11 Web Services 273

- 11.1 What is a Web Service? 274
- 11.2 Service-Oriented Architecture 277
- 11.3 JDM Web Service 278
 - 11.3.1 Overview of JDMWS Operations 279
 - 11.3.2 JDMWS Use Case 282
 - 11.3.3 JDM WSDL 288
 - 11.3.4 Data Exchange and Security in JDMWS 292
- 11.4 Enabling JDM Web Services Using JAX-RPC 293
 - 11.4.1 Overview of JAX-RPC 293
 - 11.4.2 Build JDMWS Using JAX-RPC 294
- 11.5 Summary 296
- References 297

Part III - Practice 299

Chapter 12 Practical Problem Solving 301

- 12.1 Business Scenario 1: Targeted Marketing Campaign 302
 - 12.1.1 Campaign Specifications 302
 - 12.1.2 Design of the “Campaign Optimization” Object 305
 - 12.1.3 Code Examples 306
 - 12.1.4 Scenario 1 Conclusion 320
- 12.2 Business Scenario 2: Understanding Key Factors 321
 - 12.2.1 Code Example 321
 - 12.2.2 Scenario 2 Conclusion 324
- 12.3 Business Scenario 3: Using Customer Segmentation 325
 - 12.3.1 Customer Segmentation Specifications 325
 - 12.3.2 Design of the *CustomerSegmenter* Object 327
 - 12.3.3 Code Examples 328
 - 12.3.4 Scenario 3 Conclusion 338

12.4 Summary 338

References 339

Chapter 13 Building Data Mining Tools Using JDM 341

- 13.1 Data Mining Tools 342
 - 13.1.1 Architecture of the Demonstration Interfaces 343
 - 13.1.2 Managing JDM Exceptions 345
- 13.2 Administrative Console 346
 - 13.2.1 Creating the Connection 347
 - 13.2.2 Retrieving the List of Classes That Can Be Saved 350
 - 13.2.3 Retrieving the List of Saved Objects 352
 - 13.2.4 Rename a Saved Object 355
 - 13.2.5 Delete a Saved Object from the MOR 356
- 13.3 User Interface to Build and Save a Model 356
 - 13.3.1 General Introduction 357
 - 13.3.2 Getting the Metadata 359
 - 13.3.3 Computing Statistics 361
 - 13.3.4 Retrieving the Statistics Information 364
 - 13.3.5 Saving the Physical Dataset, Build Settings, and Tasks 370
- 13.4 User Interface to Test Model Quality 376
 - 13.4.1 Getting the List of Saved Models 378
 - 13.4.2 Computing the Test Metrics 378
- 13.5 Summary 385

Chapter 14 Getting Started with JDM Web Services 387

- 14.1 A Web Service Client in PHP 387
 - 14.1.1 Filling the Input Values Using Javascript 390
 - 14.1.2 Saving the ApplySettings Object 391
 - 14.1.3 Retrieving the List of Models 394
 - 14.1.4 Executing RecordApplyTask on Models 395
- 14.2 A Web Service Client in Java 397
 - 14.2.1 How to Generate Java Classes with Axis 398
 - 14.2.2 Opening the Connection to a JDMWS Live Server 400
 - 14.2.3 Creating BuildSettings 401
 - 14.2.4 Creating a PhysicalDataSet 403
 - 14.2.5 Creating a BuildTask 404
 - 14.2.6 Executing a BuildTask 404

14.3 Summary 406
References 406

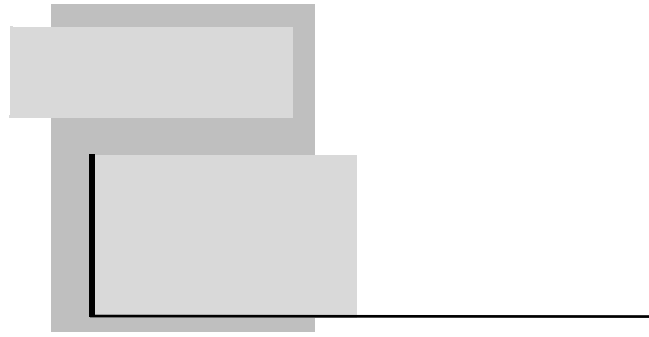
Chapter 15 The Impact of JDM on IT Infrastructure 407

15.1 What Does Data Mining Require from IT? 408
15.2 Impacts on Computing Hardware 409
15.3 Impacts on Data Storage Hardware 411
15.4 Data Access 414
 15.4.1 Data Access for Model Building 415
 15.4.2 Data Access for Apply and Test 416
15.5 Backup and Recovery 416
15.6 Scheduling 416
15.7 Workflow 417
15.8 Summary 419
References 419

Chapter 16 Vendor Implementations 421

16.1 Oracle Data Mining 421
 16.1.1 Oracle Position on JDM 422
 16.1.2 Oracle JDM Implementation Architecture 422
 16.1.3 Oracle JDM Capabilities 424
 16.1.4 Oracle JDM Extensions 425
 16.1.5 DME URI and Data URI 427
 16.1.6 Getting Started with OJDM 428
 16.1.7 Other Oracle Data Mining APIs 428
 16.1.8 Data Mining Graphical Interface Using OJDM 430
16.2 KXEN (Knowledge Extraction Engines) 431
 16.2.1 KXEN Data Mining Activity 431
 16.2.2 KXEN Position on JDM 431
 16.2.3 KXEN JDM Implementation Architecture 432
 16.2.4 KXEN JDM Capabilities 433
 16.2.5 DME URI and Data URI Specifications 435
 16.2.6 KXEN Extensions 438
 16.2.7 KXEN Web Services Implementation 439
16.3 Guidelines for New Implementers 440
 16.3.1 Standards Conformance 440
 16.3.2 Using the TCK 442
16.4 Process for New JDM Users 446
16.5 Summary 446
References 446

Part IV – Wrapping Up	449
Chapter 17 Evolution of Data Mining Standards	451
17.1 Data Mining Standards 452	
17.1.1 Predictive Model Markup Language 452	
17.1.2 Common Warehouse Metadata for Data Mining 454	
17.1.3 SQL/MM Part 6 Data Mining 455	
17.2 Java Community Process 456	
17.3 Why So Many Standards? 457	
17.4 Directions for Data Mining Standards 461	
17.5 Summary 463	
References 464	
Chapter 18 Preview of Java Data Mining 2.0	465
18.1 Transformations 466	
18.2 Time Series 469	
18.3 Apply for Association 471	
18.4 Feature Extraction 472	
18.5 Statistics 473	
18.6 Multi-target Models 474	
18.7 Text Mining 475	
18.8 Summary 476	
References 477	
Chapter 19 Summary	479
Further Reading	483
Glossary	485
Index	499
About the Authors	519



Preface

The birth of a standard is an amazing event, highlighting the ability of individuals from vastly different and often competing companies to come together to design an interface for a domain such as data mining. For JSR-73, we drew on experts from data mining tool and application vendors, as well as users of data mining technology. Data mining, as a field, is remarkably diverse in scope, encompassing capabilities from a broad range of disciplines: artificial intelligence, machine learning, statistics, data analysis, and visualization. Producing a standard in such a space is a challenging and fascinating adventure.

Within a year or so of embarking on the JDM 1.0 standard, various expert group members suggested that we'd have to write a book about Java Data Mining (JDM) someday. And indeed, here we are. Our main motivation for writing this book is to introduce data mining to a much broader audience, one that may have never used or encountered data mining before. As such, we focus less on the technical and scholarly details of data mining than on its practical understanding and application. We have tried to include a reasonably broad set of references for individuals who want to dive down to the next level of detail. However, we have strived to make data mining concepts, process, and use through JDM more accessible to Java developers, who usually do not encounter data mining, and the colleagues they will work with to develop advanced analytic applications.

Advanced analytic applications—those augmented with advanced and predictive analytics such as data mining—provide greater business intelligence, yielding insight into business problems and guidance for improved decision making. Such applications are becoming most valuable to businesses, and hence can increase revenue and profits—both for the vendors who sell them and for the businesses that use them.

Readers of this book will find a somewhat unconventional approach to data mining. Other books on data mining provide much detail on algorithms and techniques. Although this information is important to those studying machine learning or wanting to become a data analyst, other potential users of data mining are left wondering how these algorithms or techniques will be applied to solve problems. As vendors of data mining technology strive to make data mining more accessible to a broader range of users, such as business analysts, information technology (IT) specialists, and database administrators (DBAs), it is no longer the details that users require, but the big picture. Users ask, “How can I use this powerful technology to provide value within my business?” In this book, we strive to approach this and other questions from several perspectives: the software developer, the software and systems architect, and the business and data analyst. We explore these perspectives in the following section, “Guide to Readers.”

In this book, we provide insight into three key aspects of the Java Data Mining standard. The first aspect, covered in Part I, focuses on strategies for solving data mining–related business and scientific problems, and on the strategy the JDM Expert Group pursued in the design of the JDM standard. After an introduction to the data mining field, we discuss solving problems in various industries using data mining technology.

Although every industry has unique problems to solve, requiring custom and innovative solutions, each industry also shares many problems that can benefit from cross-industry solutions. For example, industries such as retail, financial services, and healthcare, as well as the public sector, all have customers. The cross-industry solution spaces include customer acquisition, customer retention, customer lifetime value, and targeted marketing.

Because data mining solutions typically do not take form or produce value in a vacuum, we then discuss the overall process, based on the industry standard data mining process CRISP-DM. Because users of data mining technology need to be minimally conversant in the terminology and concepts to problem solve with their colleagues,

we introduce the mining functions and algorithms defined in JDM. With this foundation, we explore the JDM strategy, answering questions such as: What drove the design of JDM? What is the role of standards? Lastly, before embarking on details of the Java Data Mining standard, we provide a “getting started” code example that follows the CRISP-DM process.

The second aspect, covered in Part II, focuses on the standard itself. This part introduces various concepts defined by or assimilated into the standard using examples based on business problems. After this, we explore the design of the JDM API and more detailed code examples to give readers a better understanding of how to use JDM to build applications and solve problems. Although JDM is foremost dedicated to being a standard Java language API, Java Data Mining also defines an XML schema representation for JDM objects, as well as a web services interface to enable the use of JDM functionality in a Services Oriented Architecture (SOA) environment. Part II also discusses these with specific examples of their use.

The third aspect, covered in Part III, focuses on using JDM in practice, building applications and tools that use the Java Data Mining API. We begin this part with several business scenarios (e.g., targeted marketing, key factor analysis, and customer segmentation). Because JDM is designed to be used by both application designers and data mining tool designers, we introduce code for building a simple tool graphical user interface (GUI), which manipulates JDM-persistent objects as well as enables the building and testing of a model. Having introduced web services in Part II, we give an example of a web services based application. Since data mining can impact the Information Technology (IT) infrastructure of most companies, we explore the impact of data mining along several dimensions, including hardware, software, data access, performance, and administration tools. Since the practice of using data mining often involves the use of commercial implementations, we introduce two such JDM implementations, from Oracle and KXEN. We also provide some guidelines or insights for implementers new to JDM.

Wrapping up in Part IV, we explore the evolution of data mining standards, which puts JDM in the broader context of other data mining standards. We also contrast the approaches taken by various data mining standards bodies. Since we note that no standard is ever complete, and JDM 1.1 itself covers only a subset of the possible data mining functions and algorithms, we highlight directions for JDM 2.0. We introduce features under consideration such as transformations, time series, and apply for association models, among others.

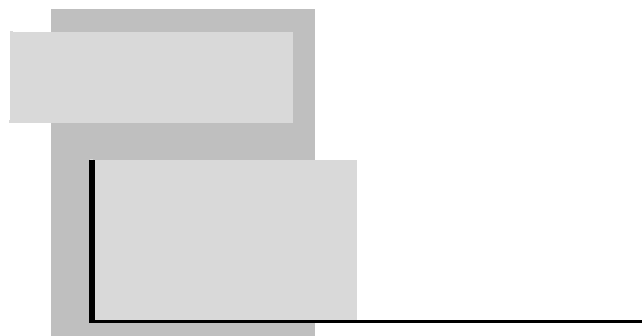
Acknowledgments

We first want to acknowledge the Java Data Mining expert group members who participated in the long process required to produce the JSR-73 standard. Their unwavering support through weekly conference calls and face-to-face meetings over the 4 years of the standards development is greatly appreciated. We also acknowledge the additional contributions of Hankil Yoon, Ka Kit Chan, Jim Dadashev, and Somesh Marepalli to the Technology Compatibility Kit (TCK) implementation, and Marwane Jai Lamimi to the Reference Implementation (RI).

We are very grateful for the general and specialist input provided by Frank Byrum, Jim Melton, and Osmar Zaiane on the developing manuscript. Over the past year, their detailed comments on both structure and content were a tremendous asset. We thank Jacek Myczkowski and Don Deutsch for their valuable comments on the final manuscript, as well as their support of the standards efforts for JSR-73 and JSR-247 at Oracle. We thank the JDM expert group members Michal Prussak, Alex Russakovsky, and Michael Smith who also provided valuable comments on the final manuscript, and David Urena and Samy Mechiri who contributed to the source code used in Part III of this book.

Of course, all remaining errors (which we expect exist despite careful review) are entirely our responsibility.

We offer our appreciation and gratitude to the wonderful people at Morgan Kaufmann Publishers as they guided us through the process of book writing and publishing. We thank Jim Melton, one of our reviewers, for putting us in contact with Diane Cerra, our talented and patient publisher, to begin this journey. We thank Diane, Asma Palmeiro, Misty Bergeron, Marilyn Rash, and Bruce Siebert who worked to make this book possible.



Guide to Readers

Data mining is becoming a mainstream technology used in business intelligence applications supporting industries such as financial services, retail, healthcare, telecommunications, and higher education, and lines of business such as marketing, manufacturing, customer experiences, customer service, and sales. Many of the business problems that data mining can solve cut across industries such as customer retention and acquisition, cross-sell, and response modeling. Due to the cost, skillsets, and complexity required to bring data mining results into an established business process, early adopters were typically big companies and research labs with correspondingly large budgets and access to statisticians and machine learning experts. In recent years, however, data mining products have simplified data mining considerably by automating the process—making the fruits of the technology more widely accessible. New algorithms and heuristics have evolved to provide good results with little or no experimentation or data preparation. In addition, the availability of data mining has increased with in-database data mining capabilities.

Java Data Mining (JDM) furthers the adoption of data mining by providing a standard Java and web services Application Programming Interface (API) for data mining. This book introduces data mining to software developers and application architects who may have heard of the benefits of data mining but are unsure how to realize these benefits. This book is also targeted at business and data analysts

who want to learn how JDM helps in developing vendor-neutral data mining solutions. It does not require a reader to be familiar with data mining, statistics, or machine learning technologies.

We have organized this book into three main parts: *strategy*, *standard*, and *practice*. In Part I, JDM Strategy, we introduce data mining in general, uses of data mining in solving industry problems, data mining processes and techniques, the role of data mining standards, and a high-level introduction to the JDM Application Programming Interface (API). Most of this part doesn't require the reader to know the Java language.

In Part II, JDM Standard, we explain the concepts used in JDM by example, explore the JDM API design and its usage, and introduce the Java Data Mining XML schema and web services. This part requires readers to know the Java language, XML, and XML schema. It gives a brief introduction to web services in Chapter 11 before discussing the JDM web services.

In Part III, JDM Practice, we illustrate practical problem solving using the JDM API. We begin by developing a sample data mining tool using JDM and a sample data mining web service using JDM. We then introduce two JDM vendor implementations, exploring their functionality, architecture, and design tradeoffs before giving some guidance to others interested in implementing a JDM-compliant system.

In Part IV, Wrapping Up, we discuss the evolution of data mining standards, where they have been and where they might go. We give a preview of some of the features proposed for JDM 2.0.

For the Software Developer



For software developers, and in particular Java and web services developers, this book introduces data mining and how to use JDM to develop data mining solutions. Part I introduces data mining and various types of business problems that can be solved using data mining, illustrates a standard process used to conduct a data mining project, describes data mining techniques used to solve business problems, explains the JDM standard strategy and why the JDM standard is necessary, and provides an overview of the JDM API. Even though software developers are not typically involved in the initial solving of a data mining problem, it is important to know

about concepts to understand the JDM API and how to develop data mining solutions.

Part II will familiarize developers with JDM concepts and the API. Readers of this part are required to know the Java language, Object Oriented Programming, the Unified Modeling Language and XML to understand the Java examples, API design concepts, JDM XML schema, and web services. This part introduces JDM concepts using examples, describes the design and usage of the Java API, and illustrates the Java Data Mining XML schema and web services interfaces.

Part III describes the use of the JDM API in practice with sample applications and detailed code examples both for the Java and web services API. It also provides JDM vendor implementation details and explains the process for other data mining vendors in adopting the JDM standard.

After reading this book, we expect the data mining knowledge gap between developers and data analysts will be greatly reduced to help them communicate more effectively when developing a data mining solution.

For the Software Architect



Data mining is often integrated with existing software applications and business processes. Understanding of data mining processes provides greater insight for architects to enable this technology in existing or new applications. For example, an architect needs to understand how data mining works to add intelligent customer offers using data mining to an existing call center application.

For architects who want to be hands-on with the JDM API (e.g., to develop prototypes), all parts of this book are useful. Part I and Part III are particularly useful for architects. Part I introduces data mining in general and provides examples of how it is currently being applied to solve business problems. Most important, it introduces the data mining process and the role of the information technology department in implementing a data mining project.

Part II will be useful to understand the API-level concepts for the architects who want to be hands-on with the API, to develop prototypes, or to mentor the developers about the use of the API.

In Part III, we provide deeper insight into how JDM can be used in practice. Chapter 16, which discusses vendor implementations, is

particularly useful for data mining software architects who are interested in developing JDM compatible API's and extensions.

After reading this book, architects should be comfortable in integrating JDM-based data mining solutions with their applications and be able to develop a strategy to operationalize data mining results with their existing applications.

For the Business/Data Analyst



For business and data analysts who want to extract actionable information from corporate data, this book provides an introduction to data mining and how it is used to solve various business problems across industries. In Part I, the data mining usage scenarios and process of implementing a data mining project will be particularly useful for the analyst unfamiliar with data mining. Chapter 5, JDM Strategy, enables analysts to understand why the JDM standard is important in implementing a data mining solution. Typically, analysts are not involved in the software implementation of the solution, yet Part II may be useful for understanding the data mining concepts used by JDM to facilitate communication with developers and data mining experts, and for using tools based on JDM.

For an analyst who is already familiar with data mining and who has expertise in data mining and statistics, this book gives details of Java Data Mining and its usage in developing data mining solutions. Data mining tools can often generate JDM-compatible code to easily deploy a solution to a JDM-compatible Data Mining Engine (DME).

After reading this book, an analyst previously unfamiliar with data mining should be able to better understand how data mining can help in solving business problems. A data mining expert analyst will be able to understand the supported data mining features in JDM and be able to communicate easily with the software architects/developers to implement a data mining solution.