INTRODUCTION TO DATA MINING

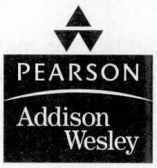
PANG-NING TAN
MICHAEL STEINBACH
VIPIN KUMAR

INTRODUCTION TO DATA MINING

PANG-NING TAN
Michigan State University

MICHAEL STEINBACH
University of Minnesota

VIPIN KUMAR
University of Minnesota
and Army High Performance
Computing Research Center



Boston San Francisco New York
London Toronto Sydney Tokyo Singapore Madrid
Mexico City Munich Paris Cape Town Hong Kong Montreal

Acquisitions Editor Matt Goldstein
Project Editor Katherine Harutunian
Production Supervisor Marilyn Lloyd
Production Services Paul C. Anagnostopoulos of Windfall Software
Marketing Manager Michelle Brown
Copyeditor Kathy Smith
Proofreader Jennifer McClain
Technical Illustration George Nichols
Cover Design Supervisor Joyce Cosentino Wells
Cover Design Night & Day Design
Cover Image © 2005 Rob Casey/Brand X Pictures
Prepress and Manufacturing Caroline Fell
Printer Hamilton Printing

Access the latest information about Addison-Wesley titles from our World Wide Web site:
http://www.aw-bc.com/computing

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress.

ISBN 0-321-32136-7

2 3 4 5 6 7 8 9 10–HAM–08 07 06 05

INTRODUCTION TO DATA MINING

To our families ...

Preface

Advances in data generation and collection are producing data sets of massive size in commerce and a variety of scientific disciplines. Data warehouses store details of the sales and operations of businesses, Earth-orbiting satellites beam high-resolution images and sensor data back to Earth, and genomics experiments generate sequence, structural, and functional data for an increasing number of organisms. The ease with which data can now be gathered and stored has created a new attitude toward data analysis: Gather whatever data you can whenever and wherever possible. It has become an article of faith that the gathered data will have value, either for the purpose that initially motivated its collection or for purposes not yet envisioned.

The field of data mining grew out of the limitations of current data analysis techniques in handling the challenges posed by these new types of data sets. Data mining does not replace other areas of data analysis, but rather takes them as the foundation for much of its work. While some areas of data mining, such as association analysis, are unique to the field, other areas, such as clustering, classification, and anomaly detection, build upon a long history of work on these topics in other fields. Indeed, the willingness of data mining researchers to draw upon existing techniques has contributed to the strength and breadth of the field, as well as to its rapid growth.

Another strength of the field has been its emphasis on collaboration with researchers in other areas. The challenges of analyzing new types of data cannot be met by simply applying data analysis techniques in isolation from those who understand the data and the domain in which it resides. Often, skill in building multidisciplinary teams has been as responsible for the success of data mining projects as the creation of new and innovative algorithms. Just as, historically, many developments in statistics were driven by the needs of agriculture, industry, medicine, and business, many of the developments in data mining are being driven by the needs of those same fields.

This book began as a set of notes and lecture slides for a data mining course that has been offered at the University of Minnesota since Spring 1998 to upper-division undergraduate and graduate students. Presentation slides

and exercises developed in these offerings grew with time and served as a basis for the book. A survey of clustering techniques in data mining, originally written in preparation for research in the area, served as a starting point for one of the chapters in the book. Over time, the clustering chapter was joined by chapters on data, classification, association analysis, and anomaly detection. The book in its current form has been class tested at the home institutions of the authors—the University of Minnesota and Michigan State University—as well as several other universities.

A number of data mining books appeared in the meantime, but were not completely satisfactory for our students—primarily graduate and undergraduate students in computer science, but including students from industry and a wide variety of other disciplines. Their mathematical and computer backgrounds varied considerably, but they shared a common goal: to learn about data mining as directly as possible in order to quickly apply it to problems in their own domains. Thus, texts with extensive mathematical or statistical prerequisites were unappealing to many of them, as were texts that required a substantial database background. The book that evolved in response to these students needs focuses as directly as possible on the key concepts of data mining by illustrating them with examples, simple descriptions of key algorithms, and exercises.

Overview Specifically, this 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, visualization, predictive modeling, association analysis, clustering, and anomaly detection. 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. In addition, this book also provides a starting point for those readers who are interested in pursuing research in data mining or related fields.

The book covers five main topics: data, classification, association analysis, clustering, and anomaly detection. Except for anomaly detection, each of these areas is covered in a pair of chapters. For classification, association analysis, and clustering, the introductory chapter covers basic concepts, representative algorithms, and evaluation techniques, while the more advanced chapter discusses advanced concepts and algorithms. The objective is to provide the reader with a sound understanding of the foundations of data mining, while still covering many important advanced topics. Because of this approach, the book is useful both as a learning tool and as a reference.

To help the readers better understand the concepts that have been presented, we provide an extensive set of examples, figures, and exercises. 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. The book also contains a comprehensive subject and author index.

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. To this end, the book was designed to be as self-contained as possible. Necessary material from statistics, linear algebra, and machine learning is either integrated into the body of the text, or for some advanced topics, covered in the appendices.

Since the chapters covering major data mining topics are self-contained, the order in which topics can be covered is quite flexible. The core material is covered in Chapters 2, 4, 6, 8, and 10. Although the introductory data chapter (2) should be covered first, the basic classification, association analysis, and clustering chapters (4, 6, and 8, respectively) can be covered in any order. Because of the relationship of anomaly detection (10) to classification (4) and clustering (8), these chapters should precede Chapter 10. Various topics can be selected from the advanced classification, association analysis, and clustering chapters (5, 7, and 9, 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 are time consuming, such hands-on assignments greatly enhance the value of the course.

Support Materials The supplements for the book are available at Addison-Wesley's Website `www.aw.com/cssupport`. Support materials available to all readers of this book include

- PowerPoint lecture slides
- Suggestions for student projects
- Data mining resources such as data mining algorithms and data sets
- On-line 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. Please contact your school's Addison-Wesley representative for information on obtaining access to this material. Comments and suggestions, as well as reports of errors, can be sent to the authors through `dmbook@cs.unm.edu`.

Acknowledgments Many people contributed to this 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, Uygar 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 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 Dan Boley, Joyce Chai, Anil Jain, Ravi

Janardan, Rong Jin, George Karypis, Haesun Park, William F. Punch, Shashi Shekhar, and Jaideep Srivastava. The collaborators on our many data mining projects, who also have our gratitude, include Ramesh Agrawal, Steve Cannon, Piet C. de Groen, Fran Hill, Yongdae Kim, Steve Klooster, Kerry Long, Nihar Mahapatra, Chris Potter, Jonathan Shapiro, Kevin Silverstein, 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, and NSF provided research support for Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. In particular, Kamal Abdali, Dick Brackney, Jagdish Chandra, Joe Coughlan, Michael Coyle, Stephen Davis, Frederica Darema, Richard Hirsch, Chandrika Kamath, Raju Namburu, N. Radhakrishnan, James Sidoran, Bhavani Thuraisingham, Walt Tiernin, Maria Zemankova, 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 Michelle Brown, Matt Goldstein, Katherine Harutunian, Marilyn Lloyd, Kathy Smith, 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: Chien-Chung Chan (University of Akron), Zhengxin Chen (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), 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).

Contents