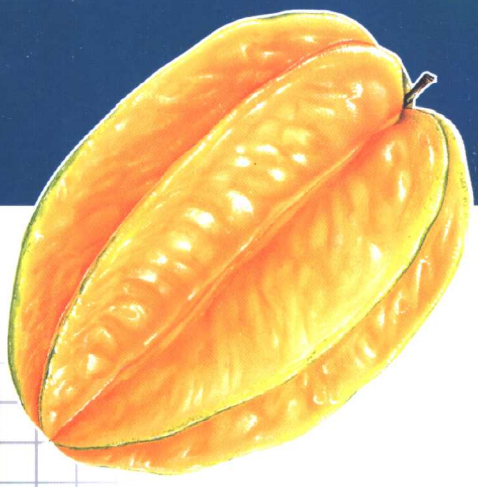
大学计算机教育国外著名教材系列（影印版）

PEARSON Addison Wesley

DATA MINING
A TUTORIAL-BASED PRIMER

数据挖掘基础教程



Richard J. Roiger
Michael W. Geatz 著

Pearson Education

清华大学出版社

大学计算机教育国外著名教材系列（影印版）

Data Mining

A Tutorial-based Primer

数据挖掘基础教程

Richard J. Roiger

Minnesota State University, Mankato

Michael W. Geatz

Information Acumen Corporation

清华大学出版社

北京

北京市版权局著作权合同登记号　图字：01-2003-7190

本书封面贴有 Pearson Education(培生教育出版集团)激光防伪标签，无标签者不得销售。

图书在版编目(CIP)数据

数据挖掘基础教程 = Data Mining: A Tutorial-based Primer /罗伊尔（Roiger, R. J.），贾茨（Geatz, M. W.）著. —影印本. —北京：清华大学出版社，2003.12
（大学计算机教育国外著名教材系列）
ISBN 7-302-07667-7

Ⅰ. 数…　Ⅱ. ①罗…　②贾…　Ⅲ. 数据采集－高等学校－教材－英文　Ⅳ. TP274

中国版本图书馆 CIP 数据核字（2003）第 106055 号

出 版 者：清华大学出版社　　**地　　址**：北京清华大学学研大厦
http://www.tup.com.cn　　**邮　　编**：100084
社总机：(010) 6277 0175　　**客户服务**：(010) 6277 6969
印 刷 者：清华大学印刷厂
装 订 者：三河市兴旺装订有限公司
发 行 者：新华书店总店北京发行所
开　　本：185×230　**印张**：25.75
版　　次：2003 年 12 月第 1 版　2003 年 12 月第 1 次印刷
书　　号：ISBN 7-302-07667-7/TP・5618
印　　数：1~5000
定　　价：43.00 元（含光盘）

出 版 说 明

进入 21 世纪，世界各国的经济、科技以及综合国力的竞争将更加激烈。竞争的中心无疑是对人才的争夺。谁拥有大量高素质的人才，谁就能在竞争中取得优势。高等教育，作为培养高素质人才的事业，必然受到高度重视。目前我国高等教育的教材更新较慢，为了加快教材的更新频率，教育部正在大力促进我国高校采用国外原版教材。

清华大学出版社从 1996 年开始，与国外著名出版公司合作，影印出版了“大学计算机教育丛书（影印版）”等一系列引进图书，受到了国内读者的欢迎和支持。跨入 21 世纪，我们本着为我国高等教育教材建设服务的初衷，在已有的基础上，进一步扩大选题内容，改变图书开本尺寸，一如既往地请有关专家挑选适用于我国高校本科及研究生计算机教育的国外经典教材或著名教材，组成本套“大学计算机教育国外著名教材系列（影印版）”，以飨读者。深切期盼读者及时将使用本系列教材的效果和意见反馈给我们。更希望国内专家、教授积极向我们推荐国外计算机教育的优秀教材，以利我们把“大学计算机教育国外著名教材系列（影印版）”做得更好，更适合高校师生的需要。

清华大学出版社
2002 年 10 月

Preface

Data mining is the process of finding useful patterns in data. The objective of data mining is to use discovered patterns to help explain current behavior or to predict future outcomes. Several aspects of the data mining process can be studied. These include:

- Data gathering and storage
- Data selection and preparation
- Model building and testing
- Interpreting and validating results
- Model application

A single book cannot concentrate on all areas of the data mining process. Although we furnish some detail about all aspects of data mining and knowledge discovery, our primary focus is centered on *model building and testing*, as well as on *interpreting and validating results*.

To help you better understand the data mining process, we provide a Microsoft Excel-based data mining tool that enables you to experimentally build and test data mining models. The Intelligent Data Analyzer (iDA), a product of Information Acumen Corporation, provides support for the business or technical analyst by offering a visual learning environment, an integrated tool set, and data mining process support. Although we recommend the provided software, you may choose to supplement this software or use an alternative software package. The parts of the text directly tied to the accompanying software are Chapters 4, 9, and Section 10 of Chapter 5.

Objectives

We wrote the text to facilitate the following student learning goals:

- Understand what data mining is and how data mining can be employed to solve real problems.
- Recognize whether a data mining solution is a feasible alternative for a specific problem.
- Step through the knowledge discovery process and write a report about the results of a data mining session.
- Apply basic statistical and nonstatistical techniques to evaluate the results of a data mining session.
- Recognize several data mining strategies and know when each strategy is appropriate.
- Develop a comprehensive understanding of how several data mining techniques build models to solve problems.
- Develop a general awareness about the structure of a data warehouse and how a data warehouse can be used to enhance business opportunities.
- Understand what on-line analytical processing (OLAP) is and how it can be applied to analyze data.
- Know that expert systems represent general models that emulate human actions.
- Know how to use a goal tree to help design a rule-based system.
- Recognize that intelligent agents are computer programs able to assist us with everyday tasks.
- Understand the types of problems that can be solved by combining an expert systems problem-solving approach and a data mining strategy.
- Know how to apply the software that accompanies this text to solve real problems.

Intended Audience

We developed most of the material for this book while teaching a one-semester introductory data mining course open to undergraduate students majoring or minoring in business or computer science. Our course also includes a unit on rule-based expert systems and intelligent agents. In writing this text, we directed our attention toward three groups of individuals:

- **Educators** who wish to teach a unit, workshop, or entire course on data mining and intelligent systems.
- **Students** who want to learn about data mining and desire hands-on experience with a data mining tool.
- **Business professionals** who need to understand how data mining and intelligent systems can be applied to help solve their business problems.

Chapter Features

We take the approach that model building is both an art and a science best understood from the perspective of learning by doing. Our view is supported by several features found within the pages of the text. The following is a partial list of these features.

- **Simple, detailed examples.** We remove much of the mystery surrounding data mining by presenting simple, detailed examples of how the various data mining techniques build their models. Because of its tutorial nature, the text is appropriate as a self-study guide as well as a college-level textbook for a course about data mining and knowledge discovery.
- **Overall tutorial style.** Selected sections in Chapters 4, 5, 6, 7, 9, and 10 offer easy to follow, step-by-step tutorials for performing data analysis.
- **Data mining sessions.** Data mining sessions allow students to work through the steps of the data mining process with the provided software. Each session is specially highlighted for easy differentiation from regular text.
- **Datasets for data mining.** A variety of datasets from business, medicine, and science are ready for data mining.
- **Aside boxes.** Aside boxes introduce the datasets for data mining and emphasize important information.
- **Web sites for data mining.** Links to several Web sites containing interesting datasets are provided.
- **Data analysis tools.** Several useful data analysis tools found within Excel are illustrated. These tools include Excel's *LINEST function* for performing linear regression analysis and *pivot tables* for summarizing and analyzing data.
- **Key term definitions.** Each chapter introduces several key terms. A list of definitions for these terms is provided at the end of each chapter.

- **End-of-chapter exercises.** The end-of-chapter exercises reinforce the techniques and concepts found within each chapter. The exercises are grouped into one of three categories—review questions, data mining questions, and computational questions. Exercises appropriate for a laboratory setting are starred.
  - *Review questions* ask basic questions about the concepts and content found within each chapter. The questions are designed to help determine if the reader understands the major points conveyed in each chapter.
  - *Data mining questions* require the reader to use one or several data mining tools to perform data mining sessions.
  - *Computational questions* have a mathematical flavor in that they require the reader to perform one or several calculations. Many of the computational questions are appropriate for challenging the more advanced student.

Chapter Content

The ordering of the chapters and the division of the book into separate parts is based on several years of experience in teaching courses on data mining and expert systems. ***Part I introduces material that is fundamental to understanding the data mining process.*** The presentation is informal and easy to follow. Basic data mining concepts, strategies, and techniques are introduced. Students learn about the types of problems that can be solved with data mining and become proficient with the software that accompanies the text. Several real-world examples of successful data mining applications are described.

Once the basic concepts are understood, ***Part II formalizes data mining problem-solving by introducing the knowledge discovery in databases (KDD) process model.*** The KDD process model is the application of the scientific method to data mining. The fact that data preprocessing is fundamental to successful data mining is emphasized. Special attention is placed on the role of the data warehouse and on data mining evaluation techniques.

Part III details several advanced data mining methods. Topics of current interest such as neural network learning, time-series analysis, logistic regression, and Web-based data mining are described. A tutorial on using the iDA neural network software is provided.

Although data mining is an appropriate solution method for many applications, there are times when this approach is not feasible. Fortunately, when data mining is not a viable choice, other options for creating useful decision-making models may be available. ***Part IV examines rule-based systems and intelligent agents as alternative methods for building models to aid in the decision-making process.*** Particular attention is directed toward combining these techniques with data mining to solve complex problems.

A brief description of the contents found within each chapter of the text follows.

Part I: Data Mining Fundamentals

- **Chapter 1** offers an overview of all aspects of the data mining process. Special emphasis is placed on helping the student determine when data mining is an appropriate problem-solving strategy.
- **Chapter 2** presents a synopsis of several common data mining strategies and techniques. Basic methods for evaluating the outcome of a data mining session are described.
- **Chapter 3** details a decision tree algorithm, the *apriori* algorithm for producing association rules, the K-Means algorithm for unsupervised clustering, and two genetic learning techniques. Tools are provided to help determine which data mining techniques should be used to solve specific problems.
- **Chapter 4** presents a tutorial introduction to the iDA software suite of data mining tools. A general methodology for performing supervised learning and unsupervised clustering is described.

Part II: Tools for Knowledge Discovery

- **Chapter 5** introduces the KDD process model as a formal methodology for solving problems with data mining. A simplified adaptation of this model is used to solve two data mining problems.
- **Chapter 6** offers a gentle introduction to data warehouse design and OLAP. A tutorial on using Excel pivot tables for data analysis is included.
- **Chapter 7** describes formal statistical and nonstatistical methods for evaluating the outcome of a data mining session. Instructions for using Excel to compute attribute correlations and display scatterplot diagrams are provided.

Part III: Advanced Data Mining Techniques

- **Chapter 8** presents two popular neural network models. A detailed explanation of neural network training is offered for the more technically inclined reader.
- **Chapter 9** offers a tutorial on applying the iDA neural network building tools to solve data mining problems. A method for using supervised learning to evaluate the results of an unsupervised neural network clustering is described.

- **Chapter 10** details several statistical techniques, including linear and logistic regression, Bayes classifier, and three unsupervised data mining methods. Instructions for using Excel's LINEST function to perform linear regression are provided.
- **Chapter 11** introduces techniques for performing time-series analysis, Web-based mining, and textual data mining. Bagging and boosting are described as methods for improving model performance.

Part IV: Intelligent Systems

The chapters of Part IV as well as appendices C, D and E are stored as Adobe PDF files on the CD that accompanies the text. To read these files you will need to have Adobe Acrobat Reader installed on you computer. To download a free copy of Adobe Acrobat Reader, visit the Web site *www.adobe.com*.

- **Chapter 12** provides an introduction to artificial intelligence and rule-based systems. A general methodology for using goal trees to build rule-based systems is described.
- **Chapter 13** reveals sources of uncertainty in rule-based systems. Fuzzy logic and Bayesian reasoning are described as methods for reasoning about uncertain information.
- **Chapter 14** introduces intelligent agents as computer programs able to assist us with everyday tasks. A model for combining intelligent agents, data mining, and expert systems to solve difficult problems is described.

Text Supplements

Each copy of this book comes with the iDA software suite of data mining tools as well as several datasets ready for data mining. Additional supplements are designed specifically for the course instructor. The following is a brief description of these supplements.

The iDA Software Package

Experiential learning is required to develop the skills required of a data mining expert. The iDA software is designed to give students this needed hands-on experience with the data mining process. The iDA software is used in several chapters to illustrate many important data mining concepts. Chapters 4, 5, 7, 9, 10, 11, and 13 have several end-of-chapter exercises designed for the iDA software.

iDA consists of a preprocessor, a report generator, and three data mining tools—ESX for supervised learning and unsupervised clustering, a neural network tool for

creating supervised backpropagation models and unsupervised self-organizing maps, and a production rule generator. As iDA is an Excel add-on, the user interface is Microsoft Excel. We chose iDA because of its flexibility and ease of use.

The iDA Dataset Package

Several datasets are included with the iDA software. The datasets come from three general application areas—business, medicine and health, and science. All datasets are in Excel format and are ready to use.

Datasets can be described along several dimensions, including the number of data instances; the number of attributes; the amount of missing or noisy data; whether data attributes are clearly defined; whether the data is categorical, numeric, or a combination of both data types; whether well-defined classes exist in the data; whether a time element is implicit in the data; whether the input attributes can differentiate between known classes contained in the data; and whether input attributes are correlated. As these factors affect the way data mining is performed, the iDA datasets were chosen to provide variety among these dimensions. The datasets also serve several general purposes. Specifically, the datasets

- Provide the beginning student with experimental data to experience the data mining process without requiring the student to deal with data preprocessing issues.
- Show the wide range of problem areas and problem types appropriate for data mining solution.
- Explain data mining outcomes.
- Illustrate the knowledge discovery process.
- Recognize that experimentation with several data mining techniques may be necessary to create a best model for a specific dataset.

The following is a short description of the datasets that are part of the iDA software package. The description includes a short statement about one or more characteristics of each dataset.

Business Applications

The Credit Card Promotion Dataset. This is a hypothetical dataset containing information about credit card holders who have accepted or rejected various promotional offerings. The dataset is used to illustrate many of the data mining techniques discussed in the text.

The Credit Card Screening Dataset. This file contains data about individuals who have applied for a credit card. The output attribute indicates

whether each individual's credit card application was accepted or rejected. The input attributes have been changed to meaningless symbols to protect confidentiality of the data.

The Deer Hunters Dataset. This dataset holds information about deer hunters who are either willing or unwilling to spend more for their next hunting trip. Several irrelevant input attributes are present in the data.

The Stock Index Dataset. The data is a time-series representation of average weekly closing prices for the Nasdaq and the Dow Jones Industrial Average.

Medicine and Health

The Cardiology Patient Dataset. This dataset holds medical information about two groups of individuals. Members of the first group have suffered one or more heart attacks. Members of the second group have not experienced a heart attack. The dataset contains a nice mix of categorical and numeric attributes.

The Spine Clinic Dataset. This dataset contains medical information about individuals who have had lower back surgery. Some of these folks have returned to work while others have not. A clear definition of the mean of each attribute is not given. The dataset contains both numeric and categorical data.

Science

The Gamma Ray Burst Dataset. This dataset contains recorded information about individual gamma-ray bursts. Gamma ray bursts are brief gamma ray flashes whose origins are outside our solar system. The bursts were observed by the Burst And Transient Source Experiment (BATSE) aboard NASA's Compton Gamma Ray Observatory between April 1991 and March 1993. Although astronomers agree that classes of gamma ray bursts exists, they do not agree on a specific class structure.

The Landsat Image Dataset. The dataset contains pixels representing a digitized satellite image of a portion of the earth's surface. Each instance has been classified into one of 15 categories. Because of the large number of individual classes, classification accuracy is affected by model-specific parameter settings.

The Temperature Dataset. This dataset offers the normal average January minimum temperature in degrees Fahrenheit for 56 U.S. cities. City latitude and longitude values are also provided. All attributes are numeric.

Miscellaneous

The Titanic Dataset. This dataset contains 2201 instances. Each instance describes attributes of an individual passenger or crew member aboard the Titanic. The output attribute indicates whether the passenger or crew member survived.

Instructor Supplements

The following supplements are provided to help the instructor organize lectures and write examinations.

- **PowerPoint slides.** Each figure and table in the text is part of a PowerPoint presentation.
- **Test questions.** Several test questions are provided for each chapter.
- **Answers to selected exercises.** Answers are given for most of the end-of-chapter exercises.
- **Lesson planner.** The lesson planner contains ideas for lecture format and points for discussion. The planner also provides suggestions for using selected end-of-chapter exercises in a laboratory setting.

Please note that these supplements are available to qualified instructors only. Contact your Addison-Wesley sales representative or an send e-mail to Computing@aw.com for access to this material.

Suggested Course Outlines

For the reader interested in the most basic understanding about the benefits and limitations of data mining, we suggest the study of Chapters 1, 2, 5, and 6. For a hands-on opportunity, include Chapter 4.

Parts I, II, and III of the text provide material for an introductory course in data mining and knowledge discovery. By including the material in Part IV, the text may also be used for a combined data mining/expert systems course that places emphasis on data mining and knowledge discovery. The prerequisite knowledge required for someone using this text is minimal. A basic understanding of spreadsheet operations, elementary statistics, and fundamental algebra is helpful.

Chapter 1 provides the essential framework for Chapters 2 through 14. Chapter 2 offers the necessary background information for Chapters 3 through 11. If you wish to provide students with an immediate hands-on learning experience, Chapter 4 can

be covered after Chapter 1 is completed. Once Chapters 2 and 3 have been studied, most of the material in Chapters 4 through 7 and 10 through 12 may be covered in any order. Chapter 9 should follow Chapter 8, and Chapters 13 and 14 should follow Chapter 12.

The text is appropriate for the undergraduate MIS or computer science student. It can also provide tutorial assistance for the graduate student who desires a working knowledge of data mining and knowledge discovery. We believe that most of the text can be covered in a single semester. Here are some options for structuring a course.

A Basic Data Mining Course for Undergraduate MIS Majors or Minors

Cover Chapters 1 through 6 in detail. However, Sections 3.3 and 3.4 of Chapter 3 may be omitted or lightly covered. Spend enough time on Chapter 4 for students to feel comfortable working with the iDA software tools.

If your students lack a course in basic statistics, Chapter 7 can be excluded or lightly covered. If Chapter 7 is skipped, spend additional time on the material in Section 2.5 (evaluating performance). Students with at least one business statistics course should be able to handle the material in Chapter 7.

Cover Chapter 8 but make Section 8.5 an optional section. Spend considerable time in Chapter 9, which shows students how to use the iDA neural net software tools.

Chapter 10 is optional. Students with some statistics in their background will find linear and logistic regression as well as Bayes classifier to be of interest. For Chapter 11, all students need some exposure to time-series analysis as well as Web-based and textual data mining. Section 11.4 is optional. As time permits, spend a day or two talking about rule-based systems (Chapter 12).

An Undergraduate MIS Course about Intelligent Systems That Emphasizes Data Mining

Follow the data mining course plan for the MIS undergraduate. Cover all material in Chapters 12 through 14. Omit the previously mentioned optional sections to make time for the added material. If time permits, supplement Chapters 12 through 14 by giving students hands-on experience with a simple rule-based expert system building tool.

A Basic Data Mining Course for Undergraduate Computer Science Majors or Minors

Cover Chapters 1 through 5 in detail. Spend a day or two on the material in Chapter 6 to provide students with a basic understanding of data warehouse design. Cover most of the material in Chapters 7 through 11. If time is an issue, you may wish to

limit your coverage of Sections 10.4, 10.5, and 11.4. Spend any extra time covering material in Chapter 12.

For a more intense course, the material on Decision Tree Attribute Selection (Appendix C) and Statistics for Performance Evaluation (Appendix D) can be covered as part of the regular course. You may wish to have students experiment with one or more of the public domain data mining tools downloadable at *www.kdnuggets.com*.

An Undergraduate Computer Science Course about Intelligent Systems That Emphasizes Data Mining

Follow the data mining course plan for the computer science undergraduate. In addition, cover the material in Chapters 12 through 14. If time is an issue, you may wish to cover only those sections of Chapters 10 and 11 that are of special interest. One plan is to cover sections 10.1, 10.2, and one subsection of 10.4.

If time permits, you can supplement Chapters 12 through 14 by giving students hands-on experience with a rule-based expert system building tool. You may also wish to have students experiment with one or more of the public domain data mining tools downloadable at *www.kdnuggets.com*.

A Data Mining Short Course

The undergraduate or graduate student interested in quickly developing a working knowledge of data mining should devote time to Chapters 1, 2, 4, and 5. A working knowledge of neural networks can be obtained through the study of Chapter 8 (Sections 8.1 through 8.4) and Chapter 9.

Acknowledgments

Many individuals helped make this book a reality. We are indebted to David Haglin and Jon Hakkila for posing critical questions on data mining performance evaluation. We are also indebted to these individuals for preprocessing several of the datasets that accompany the book. We are very grateful to Yifan Tang and Suzy for helping critique the chapters of the text. We wish to thank all of the undergraduate business and computer science students who worked with prepublished versions of our text. A special thanks goes to Information Acumen's programming team consisting of Russ Huguley, Jin Feng, and Karl Gunderson.

We also would like to thank our production coordinator Keith Henry and offer a special thanks to all of the people at Addison-Wesley for their commitment to excellence in book publishing. We are deeply indebted to our editor Maite Suarez-Rivas. Finally, we are grateful to the following reviewers of our book and found their constructive comments to be particularly helpful during revisions of the manuscript:

Ananth Grama *Purdue University*

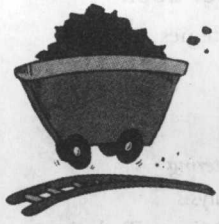
John Keane *Department of Computation, UMIST—UK*

Selwyn Piramuthu *Decision and Information Sciences, University of Florida*

Mary Ann Robbert *Bentley College*

Lynne Stokes *Southern Methodist University*

Stuart A. Varden *Pace University*



Contents