计算机科学与技术学科研究生系列教材（英文版）

PEARSON
Prentice Hall

DATA MINING
Introductory and Advanced Topics

数据挖掘教程

Margaret H. Dunham

It is the best book on data mining so far, and I would definitely adopt it for my course. Th
comprehensive and covers all of the data mining topics and algorithms of which I am aw
of coverage of each topic or method is exactly right and appropriate. Each algorithm is p
pseudocode that is sufficient for any interested readers to convert into a working implem
computer language of their choice.

— Michael H. Huhns, Univer

Discussion on distributed, parallel, and incremental algorithms is outstanding.

— Zoran Obrado

Margaret Dunham offers the experienced data base professional or graduate level Comp
student an introduction to the full spectrum of Data Mining concepts and algorithms. Usi
perspective throughout. Professor Dunham examines algorithms, data structures, data ty
complexity of algorithms and space. This text emphasizes the use of data mining concep
applications with large database components.

KEY FEATURES:

- Covers advanced topics such as Web Mining and Spatial/Temporal mining
- Includes succinct coverage of Data Warehousing, OLAP, Multidimensional Data, and P
- Provides case studies
- Offers clearly written algorithms to better understand techniques
- Includes a reference on how to use Prototypes and DM products

Pearson Education

清华大学出版社

DATA MINING
Introductory and Advanced Topics

数据挖掘教程

Margaret H. Dunham
Southern Methodist University

清华大学出版社
北京

图书在版编目(CIP)数据

数据挖掘教程 = Data Mining Introductory and Advanced Topics / [美]邓纳姆（Dunham, M. H.）著. —影印本. —北京：清华大学出版社，2003
（计算机科学与技术学科研究生系列教材）
ISBN 7-302-07383-X

Ⅰ. 数…　Ⅱ. 邓…　Ⅲ. 数据采集—研究生—教材—英文　Ⅳ. TP274

中国版本图书馆 CIP 数据核字（2003）第 091032 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
责任编辑：	张 靓	封面设计：	孟繁聪
印 刷 者：	北京密云胶印厂	装 订 者：	三河市金元装订厂
发 行 者：	新华书店总店北京发行所		
开　　本：	185×230　印张：20.75　字数：420 千字		
版　　次：	2003 年 10 月第 1 版　2003 年 10 月第 1 次印刷		
书　　号：	ISBN 7-302-07383-X/TP・5358		
印　　数：	1~5000		
定　　价：	34.80 元		

序

未来的社会是信息化的社会，计算机科学与技术在其中占据了最重要的地位，这对高素质创新型计算机人才的培养提出了迫切的要求。计算机科学与技术已经成为一门基础技术学科，理论性和技术性都很强。与传统的数学、物理和化学等基础学科相比，该学科的教育工作者既要培养学科理论研究和基本系统的开发人才，还要培养应用系统开发人才，甚至是应用人才。从层次上来讲，则需要培养系统的设计、实现、使用与维护等各个层次的人才。这就要求我们的计算机教育按照定位的需要，从知识、能力、素质三个方面进行人才培养。

硕士研究生的教育需突出"研究"，要加强理论基础的教育和科研能力的训练，使学生能够站在一定的高度去分析研究问题、解决问题。硕士研究生要通过课程的学习，进一步提高理论水平，为今后的研究和发展打下坚实的基础；通过相应的研究及学位论文撰写工作来接受全面的科研训练，了解科学研究的艰辛和科研工作者的奉献精神，培养良好的科研作风，锻炼攻关能力，养成协作精神。

高素质创新型计算机人才应具有较强的实践能力，教学与科研相结合是培养实践能力的有效途径。高水平人才的培养是通过被培养者的高水平学术成果来反映的，而高水平的学术成果主要来源于大量高水平的科研。高水平的科研还为教学活动提供了最先进的高新技术平台和创造性的工作环境，使学生得以接触最先进的计算机理论、技术和环境。高水平的科研也为高水平人才的素质教育提供了良好的物质基础。

为提高高等院校的教学质量，教育部最近实施了精品课程建设工程。由于教材是提高教学质量的关键，必须加快教材建设的步伐。为适应学科的快速发展和培养方案的需要，要采取多种措施鼓励从事前沿研究的学者参与教材的编写和更新，在教材中反映学科前沿的研究成果与发展趋势，以高水平的科研促进教材建设。同时应适当引进国外先进的原版教材，确保所有教学环节充分反映计算机学科与产业的前沿研究水平，并与未来的发展趋势相协调。

中国计算机学会教育专业委员会在清华大学出版社的大力支持下，进行了计算机科学与技术学科硕士研究生培养的系统研究。在此基础上组织来自多所全国重点大学的计算机专家和教授们编写和出版了本系列教材。作者们以自己多年来丰富的教学和科研经验为基础，认真研究和结合我国计算机科学与技术学科硕士研究生教育的特点，力图使本系列教材对我国计算机科学与技术学科硕士研究生的教学方法和教学内容的改革起到引导作用。本系列教材的系统性和理论性强，学术水平高，反映科技新发展，具有合

适的深度和广度。同时本系列教材两种语种（中文、英文）并存，三种版权（本版、外版、合作出版）形式并存，这在系列教材的出版上走出了一条新路。

相信本系列教材的出版，能够对提高我国计算机硕士研究生教材的整体水平，进而对我国大学的计算机科学与技术硕士研究生教育以及培养高素质创新型计算机人才产生积极的促进作用。

陈火旺

2003 年 9 月

陈火旺现任国防科学技术大学教授、中国工程院院士。

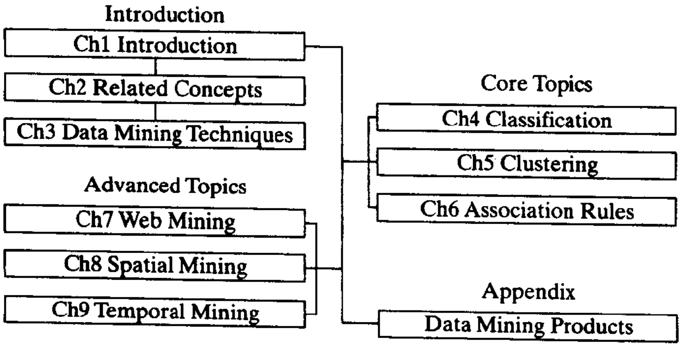
Preface

Data doubles about every year, but useful information seems to be decreasing. The area of data mining has arisen over the last decade to address this problem. It has become not only an important research area, but also one with large potential in the real world. Current business users of data mining products achieve millions of dollars a year in savings by using data mining techniques to reduce the cost of day to day business operations. Data mining techniques are proving to be extremely useful in detecting and predicting terrorism.

The purpose of this book is to introduce the reader to various data mining concepts and algorithms. The book is concise yet thorough in its coverage of the many data mining topics. Clearly written algorithms with accompanying pseudocode are used to describe approaches. A database perspective is used throughout. This means that I examine algorithms, data structures, data types, and complexity of algorithms and space. The emphasis is on the use of data mining concepts in real-world applications with large database components.

Data mining research and practice is in a state similar to that of databases in the 1960s. At that time applications programmers had to create an entire database environment each time they wrote a program. With the development of the relational data model, query processing and optimization techniques, transaction management strategies, and ad hoc query languages (SQL) and interfaces, the current environment is drastically different. The evolution of data mining techniques may take a similar path over the next few decades, making data mining techniques easier to use and develop. The objective of this book is to help in this process.

The intended audience of this book is either the experienced database professional who wishes to learn more about data mining or graduate level computer science students who have completed at least an introductory database course. The book is meant to be used as the basis of a one-semester graduate level course covering the basic data mining concepts. It may also be used as reference book for computer professionals and researchers.



The book is divided into four major parts: Introduction, Core Topics, Advanced Topics, and Appendix. The introduction covers background information needed to understand the later material. In addition, it examines topics related to data mining such as OLAP, data warehousing, information retrieval, and machine learning. In the first chapter of the introduction I provide a very cursory overview of data mining and how it relates to the complete KDD process. The second chapter surveys topics related to data mining. While this is not crucial to the coverage of data mining and need not be read to understand later chapters, it provides the interested reader with an understanding and appreciation of how data mining concepts relate to other areas. To thoroughly understand and appreciate the data mining algorithms presented in subsequent chapters, it is important that the reader realize that data mining is not an isolated subject. It has its basis in many related disciplines that are equally important on their own. The third chapter in this part surveys some techniques used to implement data mining algorithms. These include statistical techniques, neural networks, and decision trees. This part of the book provides the reader with an understanding of the basic data mining concepts. It also serves as a standalone survey of the entire data mining area.

The Core Topics covered are classification, clustering, and association rules. I view these as the major data mining functions. Other data mining concepts (such as prediction, regression, and pattern matching) may be viewed as special cases of these three. In each of these chapters I concentrate on coverage of the most commonly used algorithms of each type. Our coverage includes pseudocode for these algorithms, an explanation of them and examples illustrating their use.

The advanced topics part looks at various concepts that complicate data mining applications. I concentrate on temporal data, spatial data, and Web mining. Again, algorithms and pseudocode are provided.

In the appendix, production data mining systems are surveyed. I will keep a more up to data list on the Web page for the book. I thank all the representatives of the various companies who helped me correct and update my descriptions of their products.

All chapters include exercises covering the material in that chapter. In addition to conventional types of exercises that either test the student's understanding of the material or require him to apply what he has learned. I also include some exercises that require implementation (coding) and research. A one-semester course would cover the core topics and one or more of the advanced ones.

ACKNOWLEDGMENTS

Many people have helped with the completion of this book. Tamer Özsu provided initial advice and inspiration. My dear friend Bob Korfhage introduced me to much of computer science, including pattern matching and information retrieval. Bob, I think of you often.

I particularly thank my graduate students for contributing a great deal to some of the original wording and editing. Their assistance in reading and commenting on earlier drafts has been invaluable. Matt McBride helped me prepare most of the original slides, many of which are still available as a companion to the book. Yongqiao Xiao helped write much of the material in the Web mining chapter. He also meticulously reviewed an earlier draft of the book and corrected many mistakes. Le Gruenwald, Zahid Hossain, Yasemin Seydim, and Al Xiao performed much of the research that provided information found concerning association rules. Mario Nascimento introduced me to the world of

temporal databases, and I have used some of the information from his dissertation in the temporal mining chapter. Nat Ayewah has been very patient with his explanations of hidden Markov models and helped improve the wording of that section. Zhigang Li has introduced me to the complex world of time series and helped write the solutions manual. I've learned a lot, but still feel a novice in many of these areas.

The students in my CSE8331 class (Spring 1999, Fall 2000, and Spring 2002) at SMU have had to endure a great deal. I never realized how difficult it is to clearly word algorithm descriptions and exercises until I wrote this book. I hope they learned something even though at times the continual revisions necessary were, I'm sure, frustrating. Torsten Staab wins the prize for finding and correcting the most errors. Students in my CSE8331 class during Spring 2002 helped me prepare class notes and solutions to the exercises. I thank them for their input.

My family has been extremely supportive in this endeavor. My husband, Jim, has been (as always) understanding and patient with my odd work hours and lack of sleep. A more patient and supportive husband could not be found. My daughter Stephanie has put up with my moodiness caused by lack of sleep. Sweetie, I hope I haven't been too short-tempered with you (ILYMMMTYLM). At times I have been impatient with Kristina but you know how much I love you. My Mom, sister Martha, and brother Dave as always are there to provide support and love.

Some of the research required for this book was supported by the National Science Foundation under Grant No. IIS-9820841. I would finally like to thank the reviewers (Michael Huhns, Julia Hodger, Bob Cimikowski, Greg Speegle, Zoran Obradovic, T.Y. Lin, and James Buckly) for their many constructive comments. I tried to implement as many of these I could.

Contents

APPENDICES

PART ONE

INTRODUCTION

CHAPTER 1

Introduction

The amount of data kept in computer files and databases is growing at a phenomenal rate. At the same time, the users of these data are expecting more sophisticated information from them. A marketing manager is no longer satisfied with a simple listing of marketing contacts, but wants detailed information about customers' past purchases as well as predictions of future purchases. Simple structured/query language queries are not adequate to support these increased demands for information. Data mining steps in to solve these needs. *Data mining* is often defined as finding hidden information in a database. Alternatively, it has been called exploratory data analysis, data driven discovery, and deductive learning.

Traditional database queries (Figure 1.1), access a database using a well-defined query stated in a language such as SQL. The output of the query consists of the data from the database that satisfies the query. The output is usually a subset of the database, but it may also be an extracted view or may contain aggregations. Data mining access of a database differs from this traditional access in several ways:

- **Query:** The query might not be well formed or precisely stated. The data miner might not even be exactly sure of what he wants to see.
- **Data:** The data accessed is usually a different version from that of the original operational database. The data have been cleansed and modified to better support the mining process.
- **Output:** The output of the data mining query probably is not a subset of the database. Instead it is the output of some analysis of the contents of the database.

The current state of the art of data mining is similar to that of database query processing in the late 1960s and early 1970s. Over the next decade there undoubtedly will be great

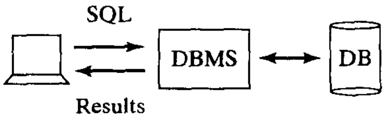


FIGURE 1.1: Database access.

strides in extending the state of the art with respect to data mining. We probably will see the development of "query processing" models, standards, and algorithms targeting the data mining applications. We probably will also see new data structures designed for the storage of databases being used for data mining applications. Although data mining is currently in its infancy, over the last decade we have seen a proliferation of mining algorithms, applications, and algorithmic approaches. Example 1.1 illustrates one such application.

EXAMPLE 1.1

Credit card companies must determine whether to authorize credit card purchases. Suppose that based on past historical information about purchases, each purchase is placed into one of four classes: (1) authorize, (2) ask for further identification before authorization, (3) do not authorize, and (4) do not authorize but contact police. The data mining functions here are twofold. First the historical data must be examined to determine how the data fit into the four classes. Then the problem is to apply this model to each new purchase. Although the second part indeed may be stated as a simple database query, the first part cannot be.

Data mining involves many different algorithms to accomplish different tasks. All of these algorithms attempt to fit a model to the data. The algorithms examine the data and determine a model that is closest to the characteristics of the data being examined. Data mining algorithms can be characterized as consisting of three parts:

- *Model*: The purpose of the algorithm is to fit a model to the data.
- *Preference*: Some criteria must be used to fit one model over another.
- *Search*: All algorithms require some technique to search the data.

In Example 1.1 the data are modeled as divided into four classes. The search requires examining past data about credit card purchases and their outcome to determine what criteria should be used to define the class structure. The preference will be given to criteria that seem to fit the data best. For example, we probably would want to authorize a credit card purchase for a small amount of money with a credit card belonging to a long-standing customer. Conversely, we would not want to authorize the use of a credit card to purchase anything if the card has been reported as stolen. The search process requires that the criteria needed to fit the data to the classes be properly defined.

As seen in Figure 1.2, the model that is created can be either predictive or descriptive in nature. In this figure, we show under each model type some of the most common data mining tasks that use that type of model.