TURING 图灵原版计算机科学系列

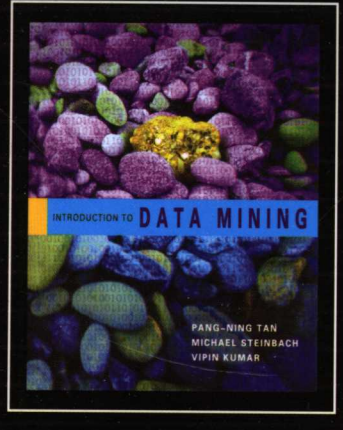
PEARSON Addison Wesley

Introduction to Data Mining

数据挖掘导论

（英文版）

[美] Pang-Ning Tan
Michael Steinbach 著
Vipin Kumar



人民邮电出版社
POSTS & TELECOM PRESS

TURING 图灵原版计算机科学系列

Introduction to Data Mining

数据挖掘导论

（英文版）

[美] Pang-Ning Tan
Michael Steinbach
Vipin Kumar 著

人民邮电出版社
POSTS & TELECOM PRESS

图书在版编目（CIP）数据

数据挖掘导论 /（美）谭等著. —北京：人民邮电出版社，2006.1
（图灵原版计算机科学系列）
ISBN 7-115-14144-4

Ⅰ. 数… Ⅱ. 谭… Ⅲ. 数据采集—英文 Ⅳ. TP274

中国版本图书馆 CIP 数据核字（2005）第 131648 号

内 容 提 要

本书对数据挖掘进行了全面介绍，旨在为读者提供将数据挖掘应用于实际问题所必需的知识。本书涵盖五个主题：数据、分类、关联分析、聚类和异常检测。除异常检测外，每个主题都有两章：前面一章讲述基本概念、代表性算法和评估技术，而后面一章较深入地讨论高级概念和算法。目的是在使读者透彻地理解数据挖掘基础的同时，还能了解更多重要的高级主题。此外，书中还提供了大量例子、图表和习题。

本书适合作为相关专业高年级本科生和研究生数据挖掘课程的教材，同时也可作为从事数据挖掘研究和应用开发工作的技术人员的参考书。

图灵原版计算机科学系列

数据挖掘导论（英文版）

◆ 著 [美] Pang-Ning Tan Michael Steinbach
Vipin Kumar

责任编辑 杨海玲

◆ 人民邮电出版社出版发行 北京市崇文区夕照寺街 14 号
邮编 100061 电子函件 315@ptpress.com.cn
网址 http://www.ptpress.com.cn
北京顺义振华印刷厂印刷
新华书店总店北京发行所经销

◆ 开本：800×1000 1/16
印张：33.25
字数：713 千字 2006 年 1 月第 1 版
印数：1－3 000 册 2006 年 1 月北京第 1 次印刷

著作权合同登记号 图字：01-2005-5235 号

ISBN 7-115-14144-4/TP・5058

定价：59.00 元

读者服务热线：(010)88593802 印装质量热线：(010)67129223

前　言

随着数据生成和收集的不断发展，在商业和科研领域产生了海量的数据集。数据仓库能够用来存储这些形形色色的数据：企业销售和运作的详细情况，绕地卫星发送回地球的高分辨率图像和遥感数据，基因组实验对越来越多的有机体产生的序列、结构和机能数据。收集和存储数据的轻松简便，已经完全改变了人们对数据分析的态度：尽可能地收集各种数据。人们开始相信收集的数据肯定会有价值，不管当初收集的目的是否明确。

数据挖掘领域兴起的根本原因，正是当前的数据分析技术在应对新的数据集所提出的挑战方面具有局限性。数据挖掘并不是要取代其他分析领域，而是将它们作为其工作的基础。尽管数据挖掘的某些主题（如关联分析）是其独有的，但是，另一些主题（如聚类、分类和异常检测）则建立在其他领域在这些主题长期工作的基础之上。事实上，数据挖掘研究者们利用已有技术的自发性已经对该领域的实力和广度以及它的快速成长贡献良多。

该领域的优势还表现在，一直强调与其他领域的研究者合作。要迎接分析新类型数据所面临的挑战，抛开理解数据的人和数据所处的领域而简单地使用数据分析技术是不可行的。通常，这种多学科研究团队既需要能够成功完成数据挖掘项目又需要能够开发新算法。正如统计学的许多发展历史上都是被农业、工业、医疗卫生和商业需求推动的一样，数据挖掘的许多发展也正在被这些领域的需求所推动。

本书源自 1998 年春季开始至今在明尼苏达大学为高年级学生和研究生开设的数据挖掘课程的讲义和教学幻灯片。在这些课程中开发的演示幻灯片和习题随着时间不断积累，成为本书的基础。数据挖掘的聚类技术综述最初是为准备该领域的研究而写的，它也成为本书一章的起点。随着时间的推移，又增加了关于数据、分类、关联分析和异常检测的几章。本书最终稿已经在作者所在的学校（明尼苏达大学和密歇根州立大学）以及其他一些大学的课堂作为教材试用了。

在此期间，出现了许多数据挖掘方面的书籍，但是都不能完全满足我们学生的需要——他们主要是计算机科学专业的研究生和本科生，也包括来自工业界和其他各学科的学生。他们的数学和计算机背景差异很大，但是都有一个共同目标：尽可能直接地学习数据挖掘，以便尽快地将其应用到各自的领域。因此，要求广泛数学和统计学预备知识的书对他们中的许多人都没有吸引力。需要坚实的数据库背景的书也有同样问题。本书根据这些学生的需求而不断发展，现在的完稿通过使用例子、关键算法的简洁描述和习题，已经尽可能直接地聚焦于数据挖掘的主要概念。

概述

具体而言，本书提供了数据挖掘的全面介绍，目的是对学生、教师、研究人员和专业人士而言容易理解和有所帮助的。本书所涵盖的领域包括数据预处理、可视化、预测建模、关联分析、聚类和异常检测。目标是讲述每个主题的基本概念和算法，从而为读者提供将数据挖掘应用于实际问题所需的必要背景。此外，本书也为有志于从事数据挖掘和相关领域研究的读者提供了很好的起点。

本书涵盖五个主题：数据、分类、关联分析、聚类和异常检测。除异常检测外，每个主题都有两章。对于分类、关联分析和聚类，前面一章讲述基本概念、代表性算法和评估技术，而后面较深入的一章讨论高级概念和算法。目的是在使读者能够透彻地理解数据挖掘基础的同时，涵盖许多重要的高级主题。由于这种安排，本书既可用作为学习工具又可用作参考书。

为了帮助读者理解书中概念，我们提供了大量例子、图表和习题。文献注释出现在每一章的结尾，是为那些对更高级的主题、重要的历史文献和当前趋势感兴趣的读者提供的。

致教师

作为一本教材，本书广泛适合于高年级本科生和研究生。对由于学习该课程的学生背景不同，可能不具有广泛的统计学和数据库知识，本书对预备知识的要求极少——不需要数据库知识，只需要适度的统计学或数学背景。本书尽可能自成一体。统计学、线性代数和机器学习的必要基础知识已经结合到正文中。

由于讨论主要数据挖掘主题的各章也是自成一体的，因此主题的讲授次序相当灵活。核心内容在第2、4、6、8和10章。尽管数据导论（第2章）应当最先讨论，但是基本的分类、关联分析和聚类（分别是第2、4、6章）可以以任意次序讲述。由于异常处理（第10章）与分类（第4章）和聚类（第8章）有一定的关系，这两章应当在第10章之前讲述。可以从高级的分类、关联分析和聚类章（分别为第5、7、9章）中挑选不同的主题，以适合课程安排和教师与学生的兴趣。我们建议教师用数据挖掘的实际项目和练习增强课程。尽管这样做很耗费时间，但是课外作业可以大大提高这门课程的价值。

支持材料

本书的辅助材料可以在Addison-Wesley的Web网站（www.aw-bc.com/cssupport）上找到。提供给所有读者的支持材料如下：

- 课程幻灯片。
- 学生项目建议。
- 数据挖掘资源，如数据挖掘算法和数据集。

- 联机指南，使用实际的数据集和数据分析软件，为本书介绍的部分数据挖掘技术提供例子讲解。

其他支持材料（包括习题答案）只向采纳本书做教材的教师提供。评论和建议以及报告错误请通过 dmbook@cs.unm.edu 发给作者。

致谢

许多人都为本书做出了贡献。我们首先向我们的家人表示感谢，这本书是献给他们的。没有他们的耐心和支持，这个项目不可能实现。

我们要感谢在明尼苏达大学和密歇根州立大学数据挖掘小组学生所做的贡献。Eui-Hong（Sam）Han 和 Mahash Joshi 为最初的数据挖掘课程提供了帮助。他们编制的某些习题和演示幻灯片已经用于本书及其辅助幻灯片中。向本书的初稿提出建议或以其他方式做出贡献的数据挖掘小组中的学生包括 Shyam Boriah、Haibin Cheng、Varun Chandola、Eric Eilertson、Levent Ertőz、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 和 Pusheng Zhang。我们还要感谢明尼苏达大学和密歇根州立大学选修数据挖掘课程的学生，他们使用了本书的早期书稿，并提供了极富价值的反馈。我们特别感谢 Bernardo Craemer、Arifin Ruslim、Jamshid Vayghan 和 Yu Wei 的有益的建议。

Joydeep Ghosh（得克萨斯大学）和 Sanjay Ranka（佛罗里达大学）试用了本书的早期版本。我们也直接从得克萨斯大学下列学生那里获得了许多有用的建议：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 和 Mei Yang。

Ronald Kostoff（ONR）阅读了聚类一章的早期版本，并提出了许多建议。Musetta Steinbach 帮助发现了图中的错误。

我们要感谢明尼苏达大学和密歇根州立大学的同事，他们帮助创建了良好的数据挖掘研究环境。他们是 Dan Boley、Joyce Chai、Anil Jain、Ravi Janardan、Rong Jin、Goerge Karypis、Haesun Park、William F. Punch、Shashi Shekhar 和 Jaideep Srivastava。我们还要向我们的数据挖掘项目的合作者表示谢意，他们是 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 和 Zhi-Li Zhang。

明尼苏达大学和密歇根州立大学计算机科学系为该项目提供了计算资源和支持环境。ARDA、ARL、ARO、DOE、NASA 和 NSF 为本书作者提供了研究资助。特别应该提到的是，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 和 Xiaodong Zhang 有力地支持了我们的数据挖掘和高性能计算研究。

与 Pearson Education 工作人员的合作令人愉快。具体地，我们要感谢 Michelle Brown、Matt Goldstein、Katherine Harutunian、Marilyn Lloyd、Kathy Smith 和 Joyce Wells。我们还要感谢 George Nichols 帮助绘图，Paul Anagnostopoulos 提供 LaTeX 支持。我们感谢 Pearson 邀请的审稿人：Chien-Chung Chan（阿克伦大学）、Zhengxin Chen（内布拉斯加大学奥马哈分校）、Chris Clifton（普度大学）、Joydeep Ghosh（得克萨斯大学奥斯汀分校）、Nazli Goharian（伊利诺伊理工学院）、J. Michael Hardin（阿拉巴马大学）、James Hearne（西华盛顿大学），Hillol Kargupta（马里兰大学巴尔的摩县分校和 Agnik 公司）、Eamonn Keogh（加利福尼亚大学埃尔文分校）、Bing Liu（伊利诺伊大学芝加哥分校）、Mariofanna Milanova（阿肯色大学小石城分校）、Srinivasan Parthasarathy（俄亥俄州立大学）、Zbigniew W. Ras（北卡罗莱纳大学夏洛特分校）、Xintao Wu（北卡罗莱纳大学夏洛特分校）和 Mohammed J. Zaki（伦斯勒理工学院）。

Contents

1 Introduction 1

1.1 What Is Data Mining? 2
1.2 Motivating Challenges 3
1.3 The Origins of Data Mining 4
1.4 Data Mining Tasks 5
1.5 Scope and Organization of the Book 8
1.6 Bibliographic Notes 9
1.7 Exercises 12

2 Data 13

2.1 Types of Data 15
2.1.1 Attributes and Measurement 15
2.1.2 Types of Data Sets 20
2.2 Data Quality 25
2.2.1 Measurement and Data Collection Issues 26
2.2.2 Issues Related to Applications 31
2.3 Data Preprocessing 32
2.3.1 Aggregation 32
2.3.2 Sampling 34
2.3.3 Dimensionality Reduction 36
2.3.4 Feature Subset Selection 37
2.3.5 Feature Creation 39
2.3.6 Discretization and Binarization 41
2.3.7 Variable Transformation 45
2.4 Measures of Similarity and Dissimilarity 47
2.4.1 Basics 47
2.4.2 Similarity and Dissimilarity between Simple Attributes 49
2.4.3 Dissimilarities between Data Objects 50
2.4.4 Similarities between Data Objects 52
2.4.5 Examples of Proximity Measures 53
2.4.6 Issues in Proximity Calculation 58
2.4.7 Selecting the Right Proximity Measure 60
2.5 Bibliographic Notes 61
2.6 Exercises 64

3 Exploring Data 71
3.1 The Iris Data Set ... 71
3.2 Summary Statistics ... 72
3.2.1 Frequencies and the Mode ... 72
3.2.2 Percentiles ... 73
3.2.3 Measures of Location: Mean and Median ... 73
3.2.4 Measures of Spread: Range and Variance ... 75
3.2.5 Multivariate Summary Statistics ... 76
3.2.6 Other Ways to Summarize the Data ... 77
3.3 Visualization ... 77
3.3.1 Motivations for Visualization ... 77
3.3.2 General Concepts ... 78
3.3.3 Techniques ... 81
3.3.4 Visualizing Higher-Dimensional Data ... 90
3.3.5 Do's and Don'ts ... 94
3.4 OLAP and Multidimensional Data Analysis ... 95
3.4.1 Representing Iris Data as a Multidimensional Array ... 95
3.4.2 Multidimensional Data: The General Case ... 97
3.4.3 Analyzing Multidimensional Data ... 98
3.4.4 Final Comments on Multidimensional Data Analysis ... 101
3.5 Bibliographic Notes ... 102
3.6 Exercises ... 103

4 Classification: Basic Concepts, Decision Trees, and Model Evaluation 105
4.1 Preliminaries ... 105
4.2 General Approach to Solving a Classification Problem ... 107
4.3 Decision Tree Induction ... 108
4.3.1 How a Decision Tree Works ... 108
4.3.2 How to Build a Decision Tree ... 110
4.3.3 Methods for Expressing Attribute Test Conditions ... 112
4.3.4 Measures for Selecting the Best Split ... 114
4.3.5 Algorithm for Decision Tree Induction ... 119
4.3.6 An Example: Web Robot Detection ... 120
4.3.7 Characteristics of Decision Tree Induction ... 122
4.4 Model Overfitting ... 125
4.4.1 Overfitting Due to Presence of Noise ... 127
4.4.2 Overfitting Due to Lack of Representative Samples ... 129
4.4.3 Overfitting and the Multiple Comparison Procedure ... 129
4.4.4 Estimation of Generalization Errors ... 131
4.4.5 Handling Overfitting in Decision Tree Induction ... 134

4.5 Evaluating the Performance of a Classifier 135
4.5.1 Holdout Method 136
4.5.2 Random Subsampling 136
4.5.3 Cross-Validation 136
4.5.4 Bootstrap 137
4.6 Methods for Comparing Classifiers 137
4.6.1 Estimating a Confidence Interval for Accuracy 138
4.6.2 Comparing the Performance of Two Models 139
4.6.3 Comparing the Performance of Two Classifiers 140
4.7 Bibliographic Notes 141
4.8 Exercises 144

5 Classification: Alternative Techniques **151**
5.1 Rule-Based Classifier 151
5.1.1 How a Rule-Based Classifier Works 153
5.1.2 Rule-Ordering Schemes 154
5.1.3 How to Build a Rule-Based Classifier 155
5.1.4 Direct Methods for Rule Extraction 155
5.1.5 Indirect Methods for Rule Extraction 161
5.1.6 Characteristics of Rule-Based Classifiers 163
5.2 Nearest-Neighbor classifiers 163
5.2.1 Algorithm 165
5.2.2 Characteristics of Nearest-Neighbor Classifiers 165
5.3 Bayesian Classifiers 166
5.3.1 Bayes Theorem 166
5.3.2 Using the Bayes Theorem for Classification 168
5.3.3 Naïve Bayes Classifier 169
5.3.4 Bayes Error Rate 175
5.3.5 Bayesian Belief Networks 176
5.4 Artificial Neural Network (ANN) 181
5.4.1 Perceptron 181
5.4.2 Multilayer Artificial Neural Network 184
5.4.3 Characteristics of ANN 187
5.5 Support Vector Machine (SVM) 188
5.5.1 Maximum Margin Hyperplanes 188
5.5.2 Linear SVM: Separable Case 190
5.5.3 Linear SVM: Nonseparable Case 195
5.5.4 Nonlinear SVM 198
5.5.5 Characteristics of SVM 203
5.6 Ensemble Methods 203
5.6.1 Rationale for Ensemble Method 203

5.6.2 Methods for Constructing an Ensemble Classifier 204
5.6.3 Bias-Variance Decomposition 206
5.6.4 Bagging 209
5.6.5 Boosting 211
5.6.6 Random Forests 215
5.6.7 Empirical Comparison among Ensemble Methods 216
5.7 Class Imbalance Problem 217
5.7.1 Alternative Metrics 218
5.7.2 The Receiver Operating Characteristic Curve 220
5.7.3 Cost-Sensitive Learning 223
5.7.4 Sampling-Based Approaches 225
5.8 Multiclass Problem 226
5.9 Bibliographic Notes 228
5.10 Exercises 233

6 Association Analysis: Basic Concepts and Algorithms **241**
6.1 Problem Definition 242
6.2 Frequent Itemset Generation 244
6.2.1 The *Apriori* Principle 246
6.2.2 Frequent Itemset Generation in the *Apriori* Algorithm 247
6.2.3 Candidate Generation and Pruning 249
6.2.4 Support Counting 252
6.2.5 Computational Complexity 255
6.3 Rule Generation 257
6.3.1 Confidence-Based Pruning 258
6.3.2 Rule Generation in *Apriori* Algorithm 258
6.3.3 An Example: Congressional Voting Records 259
6.4 Compact Representation of Frequent Itemsets 260
6.4.1 Maximal Frequent Itemsets 260
6.4.2 Closed Frequent Itemsets 262
6.5 Alternative Methods for Generating Frequent Itemsets 264
6.6 FP-Growth Algorithm 268
6.6.1 FP-Tree Representation 268
6.6.2 Frequent Itemset Generation in FP-Growth Algorithm 270
6.7 Evaluation of Association Patterns 273
6.7.1 Objective Measures of Interestingness 274
6.7.2 Measures beyond Pairs of Binary Variables 282
6.7.3 Simpson's Paradox 283
6.8 Effect of Skewed Support Distribution 285
6.9 Bibliographic Notes 288
6.10 Exercises 298

7 Association Analysis: Advanced Concepts **307**
7.1 Handling Categorical Attributes 307
7.2 Handling Continuous Attributes 309
7.2.1 Discretization-Based Methods 310
7.2.2 Statistics-Based Methods 312
7.2.3 Non-discretization Methods 314
7.3 Handling a Concept Hierarchy 316
7.4 Sequential Patterns 318
7.4.1 Problem Formulation 318
7.4.2 Sequential Pattern Discovery 320
7.4.3 Timing Constraints 323
7.4.4 Alternative Counting Schemes 327
7.5 Subgraph Patterns 328
7.5.1 Graphs and Subgraphs 329
7.5.2 Frequent Subgraph Mining 330
7.5.3 *Apriori*-like Method 332
7.5.4 Candidate Generation 333
7.5.5 Candidate Pruning 338
7.5.6 Support Counting 340
7.6 Infrequent Patterns 340
7.6.1 Negative Patterns 341
7.6.2 Negatively Correlated Patterns 342
7.6.3 Comparisons among Infrequent Patterns, Negative Patterns, and Negatively Correlated Patterns 343
7.6.4 Techniques for Mining Interesting Infrequent Patterns 344
7.6.5 Techniques Based on Mining Negative Patterns 345
7.6.6 Techniques Based on Support Expectation 347
7.7 Bibliographic Notes 350
7.8 Exercises 353

8 Cluster Analysis: Basic Concepts and Algorithms **363**
8.1 Overview 365
8.1.1 What Is Cluster Analysis? 365
8.1.2 Different Types of Clusterings 366
8.1.3 Different Types of Clusters 368
8.2 K–means 370
8.2.1 The Basic K-means Algorithm 371
8.2.2 K-means: Additional Issues 378
8.2.3 Bisecting K-means 380
8.2.4 K-means and Different Types of Clusters 381

8.2.5 Strengths and Weaknesses 383
8.2.6 K-means as an Optimization Problem 383
8.3 Agglomerative Hierarchical Clustering 385
8.3.1 Basic Agglomerative Hierarchical Clustering Algorithm 385
8.3.2 Specific Techniques 387
8.3.3 The Lance-Williams Formula for Cluster Proximity 391
8.3.4 Key Issues in Hierarchical Clustering 391
8.3.5 Strengths and Weaknesses 393
8.4 DBSCAN 393
8.4.1 Traditional Density: Center-Based Approach 393
8.4.2 The DBSCAN Algorithm 394
8.4.3 Strengths and Weaknesses 398
8.5 Cluster Evaluation 398
8.5.1 Overview 399
8.5.2 Unsupervised Cluster Evaluation Using Cohesion and Separation 401
8.5.3 Unsupervised Cluster Evaluation Using the Proximity Matrix 406
8.5.4 Unsupervised Evaluation of Hierarchical Clustering 408
8.5.5 Determining the Correct Number of Clusters 409
8.5.6 Clustering Tendency 410
8.5.7 Supervised Measures of Cluster Validity 411
8.5.8 Assessing the Significance of Cluster Validity Measures 414
8.6 Bibliographic Notes 416
8.7 Exercises 419

9 Cluster Analysis: Additional Issues and Algorithms 427
9.1 Characteristics of Data, Clusters, and Clustering Algorithms 427
9.1.1 Example: Comparing K-means and DBSCAN 428
9.1.2 Data Characteristics 429
9.1.3 Cluster Characteristics 430
9.1.4 General Characteristics of Clustering Algorithms 431
9.2 Prototype-Based Clustering 433
9.2.1 Fuzzy Clustering 433
9.2.2 Clustering Using Mixture Models 437
9.2.3 Self-Organizing Maps (SOM) 446
9.3 Density-Based Clustering 451
9.3.1 Grid-Based Clustering 451
9.3.2 Subspace Clustering 454
9.3.3 DENCLUE: A Kernel-Based Scheme for Density-Based Clustering 457
9.4 Graph-Based Clustering 460
9.4.1 Sparsification 461

9.4.2 Minimum Spanning Tree (MST) Clustering 462
9.4.3 OPOSSUM: Optimal Partitioning of Sparse Similarities Using METIS 463
9.4.4 Chameleon: Hierarchical Clustering with Dynamic Modeling 464
9.4.5 Shared Nearest Neighbor Similarity 468
9.4.6 The Jarvis-Patrick Clustering Algorithm 471
9.4.7 SNN Density 472
9.4.8 SNN Density-Based Clustering 473
9.5 Scalable Clustering Algorithms 475
9.5.1 Scalability: General Issues and Approaches 476
9.5.2 BIRCH 477
9.5.3 CURE 479
9.6 Which Clustering Algorithm? 482
9.7 Bibliographic Notes 484
9.8 Exercises 488

10 Anomaly Detection **491**
10.1 Preliminaries 492
10.1.1 Causes of Anomalies 492
10.1.2 Approaches to Anomaly Detection 493
10.1.3 The Use of Class Labels 494
10.1.4 Issues 495
10.2 Statistical Approaches 496
10.2.1 Detecting Outliers in a Univariate Normal Distribution 497
10.2.2 Outliers in a Multivariate Normal Distribution 499
10.2.3 A Mixture Model Approach for Anomaly Detection 500
10.2.4 Strengths and Weaknesses 502
10.3 Proximity-Based Outlier Detection 502
10.3.1 Strengths and Weaknesses 503
10.4 Density-Based Outlier Detection 504
10.4.1 Detection of Outliers Using Relative Density 505
10.4.2 Strengths and Weaknesses 506
10.5 Clustering-Based Techniques 506
10.5.1 Assessing the Extent to Which an Object Belongs to a Cluster 507
10.5.2 Impact of Outliers on the Initial Clustering 509
10.5.3 The Number of Clusters to Use 509
10.5.4 Strengths and Weaknesses 509
10.6 Bibliographic Notes 510
10.7 Exercises 513

1

Introduction

Rapid advances in data collection and storage technology have enabled organizations to accumulate vast amounts of data. However, extracting useful information has proven extremely challenging. Often, traditional data analysis tools and techniques cannot be used because of the massive size of a data set. Sometimes, the non-traditional nature of the data means that traditional approaches cannot be applied even if the data set is relatively small. In other situations, the questions that need to be answered cannot be addressed using existing data analysis techniques, and thus, new methods need to be developed.

Data mining is a technology that blends traditional data analysis methods with sophisticated algorithms for processing large volumes of data. It has also opened up exciting opportunities for exploring and analyzing new types of data and for analyzing old types of data in new ways. In this introductory chapter, we present an overview of data mining and outline the key topics to be covered in this book. We start with a description of some well-known applications that require new techniques for data analysis.

Business Point-of-sale data collection (bar code scanners, radio frequency identification (RFID), and smart card technology) have allowed retailers to collect up-to-the-minute data about customer purchases at the checkout counters of their stores. Retailers can utilize this information, along with other business-critical data such as Web logs from e-commerce Web sites and customer service records from call centers, to help them better understand the needs of their customers and make more informed business decisions.

Data mining techniques can be used to support a wide range of business intelligence applications such as customer profiling, targeted marketing, workflow management, store layout, and fraud detection. It can also help retailers answer important business questions such as "Who are the most profitable customers?" "What products can be cross-sold or up-sold?" and "What is the revenue outlook of the company for next year?" Some of these questions motivated the creation of association analysis (Chapters 6 and 7), a new data analysis technique.

Medicine, Science, and Engineering Researchers in medicine, science, and engineering are rapidly accumulating data that is key to important new discoveries. For example, as an important step toward improving our understanding of the Earth's climate system, NASA has deployed a series of Earth-orbiting satellites that continuously generate global observations of the land surface, oceans, and atmosphere. However, because of the size and spatiotemporal nature of the data, traditional methods are often not suitable for analyzing these data sets. Techniques developed in data mining can aid Earth scientists in answering questions such as "What is the relationship between the frequency and intensity of eco-

system disturbances such as droughts and hurricanes to global warming?" "How is land surface precipitation and temperature affected by ocean surface temperature?" and "How well can we predict the beginning and end of the growing season for a region?"

As another example, researchers in molecular biology hope to use the large amounts of genomic data currently being gathered to better understand the structure and function of genes. In the past, traditional methods in molecular biology allowed scientists to study only a few genes at a time in a given experiment. Recent breakthroughs in microarray technology have enabled scientists to compare the behavior of thousands of genes under various situations. Such comparisons can help determine the function of each gene and perhaps isolate the genes responsible for certain diseases. However, the noisy and high-dimensional nature of data requires new types of data analysis. In addition to analyzing gene array data, data mining can also be used to address other important biological challenges such as protein structure prediction, multiple sequence alignment, the modeling of biochemical pathways, and phylogenetics.

1.1 What Is Data Mining?

Data mining is the process of automatically discovering useful information in large data repositories. Data mining techniques are deployed to scour large databases in order to find novel and useful patterns that might otherwise remain unknown. They also provide capabilities to predict the outcome of a future observation, such as predicting whether a newly arrived customer will spend more than $100 at a department store.

Not all information discovery tasks are considered to be data mining. For example, looking up individual records using a database management system or finding particular Web pages via a query to an Internet search engine are tasks related to the area of **information retrieval**. Although such tasks are important and may involve the use of the sophisticated algorithms and data structures, they rely on traditional computer science techniques and obvious features of the data to create index structures for efficiently organizing and retrieving information. Nonetheless, data mining techniques have been used to enhance information retrieval systems.

Data Mining and Knowledge Discovery

Data mining is an integral part of **knowledge discovery in databases (KDD)**, which is the overall process of converting raw data into useful information, as shown in Figure 1.1. This process consists of a series of transformation steps, from data preprocessing to postprocessing of data mining results.

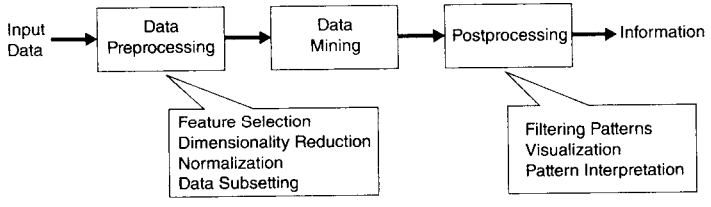


Figure 1.1. The process of knowledge discovery in databases (KDD).

The input data can be stored in a variety of formats (flat files, spreadsheets, or relational tables) and may reside in a centralized data repository or be distributed across multiple sites. The purpose of **preprocessing** is to transform the raw input data into an appropriate format for subsequent analysis. The steps involved in data preprocessing include fusing data from multiple sources, cleaning data to remove noise and duplicate observations, and selecting records and features that are relevant to the data mining task at hand. Because of the many ways data can be collected and stored, data preprocessing is perhaps the most laborious and time-consuming step in the overall knowledge discovery process.

"Closing the loop" is the phrase often used to refer to the process of integrating data mining results into decision support systems. For example, in business applications, the insights offered by data mining results can be integrated with campaign management tools so that effective marketing promotions can be conducted and tested. Such integration requires a **postprocessing** step that ensures that only valid and useful results are incorporated into the decision support system. An example of postprocessing is visualization (see Chapter 3), which allows analysts to explore the data and the data mining results from a variety of viewpoints. Statistical measures or hypothesis testing methods can also be applied during postprocessing to eliminate spurious data mining results.

1.2 Motivating Challenges

As mentioned earlier, traditional data analysis techniques have often encountered practical difficulties in meeting the challenges posed by new data sets. The following are some of the specific challenges that motivated the development of data mining.

Scalability Because of advances in data generation and collection, data sets with sizes of gigabytes, terabytes, or even petabytes are becoming common. If data mining algorithms are to handle these massive data sets, then they must be scalable. Many data mining algorithms employ special search strategies to handle exponential search problems. Scalability may also require the implementation of novel data structures to access individual records in an efficient manner. For instance, out-of-core algorithms may be necessary when processing data sets that cannot fit into main memory. Scalability can also be improved by using sampling or developing parallel and distributed algorithms.

High Dimensionality It is now common to encounter data sets with hundreds or thousands of attributes instead of the handful common a few decades ago. In bioinformatics, progress in microarray technology has produced gene expression data involving thousands of features. Data sets with temporal or spatial components also tend to have high dimensionality. For example, consider a data set that contains measurements of temperature at various locations. If the temperature measurements are taken repeatedly for an extended period, the number of dimensions (features) increases in proportion to the number of measurements taken. Traditional data analysis techniques that were developed for low-dimensional data often do not work well for such high-dimensional data. Also, for some data analysis algorithms, the computational complexity increases rapidly as the dimensionality (the number of features) increases.

Heterogeneous and Complex Data Traditional data analysis methods often deal with