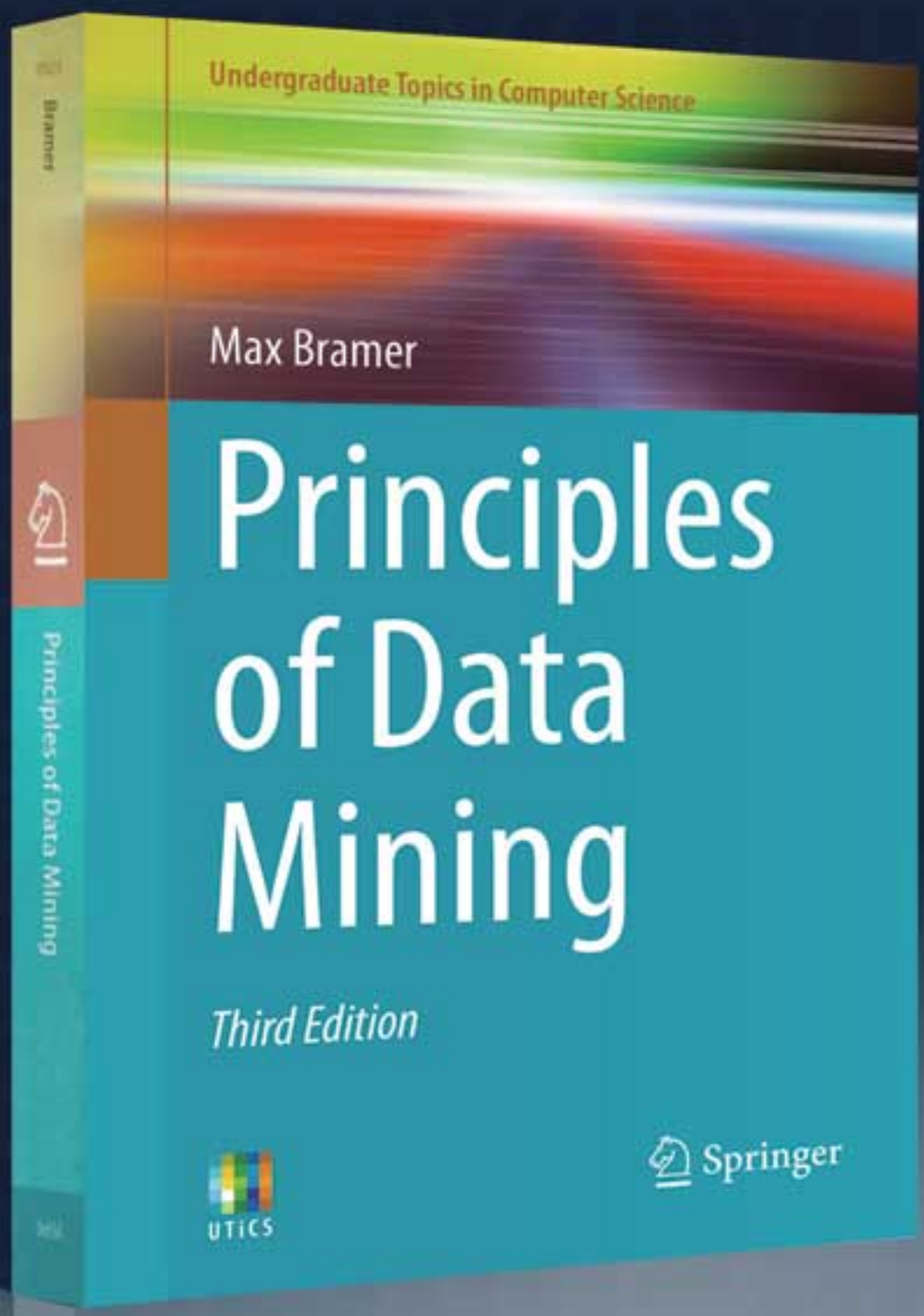
国外计算机科学经典教材

Principles of Data Mining, Third Edition

数据挖掘原理

(第3版)

[英] 麦克斯•布拉默(Max Bramer) 著
王 净 译



Springer
清华大学出版社

国外计算机科学经典教材

数据挖掘原理

(第3版)

[英] 麦克斯·布拉默(Max Bramer) 著

王 净 译

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：01-2018-7422

图书在版编目(CIP)数据

数据挖掘原理 / (英)麦克斯 • 布拉默(Max Bramer) 著；王净 译. —3版. —北京：清华大学出版社，2019
(国外计算机科学经典教材)
书名原文：Principles of Data Mining, Third Edition
ISBN 978-7-302-52681-0

Ⅰ. ①数…　Ⅱ. ①麦… ②王…　Ⅲ. ①数据采集—高等学校—教材　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2019)第 057435 号

责任编辑：王　军　韩宏志
装帧设计：孔祥峰
责任校对：成凤进
责任印制：李红英

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市少明印务有限公司
经　　销：全国新华书店
开　　本：170mm×240mm　　**印　　张**：27.5　　**字　　数**：537 千字
版　　次：2019 年 9 月第 1 版　　**印　　次**：2019 年 9 月第 1 次印刷
定　　价：79.80 元

产品编号：081479-01

译 者 序

数据挖掘(Data Mining)是从不完全的、有噪声的、模糊的大量随机数据中提取出隐含的、人们事先不知道但潜在有用的信息和知识的过程。随着信息技术的高速发展，商业和科研领域积累的数据量急剧增长，动辄以 TB 计，如何从海量数据中提取有用信息成为当务之急。为顺应这种趋势，数据挖掘技术应运而生。数据挖掘是知识发现的关键。

本书共分 22 章，另有 5 个附录，系统介绍数据挖掘与应用的基本算法，探讨每种算法的基本原理，并列举大量示例加以演示，这种理论与实践相结合的方式极大地方便了读者对抽象算法的理解和掌握。第 1~3 章主要介绍数据挖掘的基本概念和分类，重点分析朴素贝叶斯算法和最近邻算法。第 4~15 章是全书的重点，浓墨重彩地描述决策树，包括属性选择的不同标准、预测分类器的精度、避免过度拟合、连续属性离散化、熵、归纳分类的模块化规则、集成分类等。第 16~18 章主要介绍关联规则挖掘的相关内容。第 19 章和第 20 章讲述聚类和文本挖掘。第 21 章和第 22 章讲述如何对流数据进行分类。为巩固学习效果，每章结尾处提供若干自我评估练习，以便读者自查，附录 E 给出练习题答案。此外，与其他许多书籍不同，在学习本书的过程中，不需要太多数学知识即可理解相关内容。附录 A 介绍相关数学内容和符号，附录 B 详细列出书中所用的数据集，附录 C 提供其他一些关于数据挖掘的资源，附录 D 列出术语表。

本书可作为本科生或硕士研究生的数据挖掘教科书，适用的学科广泛，包括计算机科学、商业研究、市场营销、人工智能、生物信息学和法医学等。对于那些希望进一步提高自身能力的技术或管理人员来说，本书也是一本优秀的自学书籍。

参与本书翻译的人员有王净、范园芳、胡训强和晏峰，最终由王净负责统稿。

译者在翻译过程中，尽量保持原书特色，并对书中出现的术语和难句进行了认真推敲和研究。但毕竟有少量技术是译者在自己的研究领域中不曾遇到过的，所以疏漏和争议之处在所难免，望广大读者提出宝贵意见。

最后，希望广大读者能多花些时间细细品味这本凝聚了作者和译者大量心血的书籍，为自己将来的职业生涯奠定良好基础。

王净

作于广州

致　　谢

首先感谢我的女儿 Bryony，她帮我绘制了许多复杂的图表并提出设计建议。其次感谢 Frederic Stahl 博士，他就第 21 章和第 22 章给出了许多宝贵建议，最后要感谢我的妻子 Dawn，她对本书草案给出了相当宝贵的意见。不过，最终版本中的任何错误仍然由我负责。

Max Bramer

关于第3版的说明

自第 1 版以来，可用于数据挖掘的数据量大幅增加。第 1 章所引用的数字现在看起来相当保守。根据 IBM 于 2016 年所做的统计，每天从各种传感器、移动设备、在线交易和社交网络生成的数据量高达 2.5YB，仅过去两年就创建了世界上 90%的数据。现在，每天的数据流记录超过 100 万条。因此，第 3 版新增了两章，用于详细解释对流数据进行分类的算法。

前　言

本书面向计算机科学、商业研究、市场营销、人工智能、生物信息学和法医学专业的学生，可用作本科生或硕士研究生的入门教材。同时，对于那些希望进一步提高自身能力的技术或管理人员来说，本书也是一本极佳的自学书籍。本书所涉及的内容远超一般的数据挖掘入门书籍。与许多其他书籍不同的是，在学习过程中你不需要拥有太多的数学知识即可理解相关内容。

数学是一种可以表达复杂思想的语言。遗憾的是，99%的人都无法很好地掌握这门语言；很多人很早就开始在学校学习一些基础知识，但学习过程往往充满曲折。

本书涉及数学公式较少，将重点介绍相关概念。但遗憾的是，完全不使用数学符号是不可能的。附录 A 给出开始学习本书需要掌握的所有内容。对于那些在学校学习数学的人来说，这些内容应该是非常熟悉的。掌握这些内容后，其他内容就较好理解了。如果觉得某些数学符号难以理解，通常可放心地忽略它们，只需要关注结果和给出的详细示例即可。而对于那些希望更深入理解数据挖掘的数学基础知识的人来说，可参考附录 C 中列出的内容。

过去，没有一本关于数据挖掘的入门书可使你具备该领域的研究水平——但现在，这样的日子已经过去了。本书的重点是介绍基本技术，而不是展示当今最新的数据挖掘技术，因为大多数情况下，当拿到一本书时，书中介绍的技术可能已被其他更新的技术取代了。一旦掌握了基本技术，你可通过多种渠道来了解该领域的最新进展。附录 C 列出一些常用资源，而其他附录包括有关本书示例中使用的主要数据集的信息，供你在自己的项目中使用。此外附录 D 包括技术术语表。

为便于检查对所学知识的掌握情况，每章都包含自我评估练习。参考答案见附录 E。

另外说明一下，本书涉及大量数据集、属性和值，也涉及不少数学公式，字母繁多，格式复杂。为保证全书的科学性和严谨性，中文书中，字母的正斜体与英文原书基本保持统一。

书末列出全书各章正文中引用的参考文献。读者在阅读正文时，会不时看到引用；引用的形式为[*]，其中*为数字编号。遇到此类引用时，读者可跳转到书末，查阅相关信息。

目　录

第 1 章

数据挖掘简介

1.1 数据爆炸

现代计算机系统以令人难以置信的速度从各种来源收集数据。从街头的 POS 机到用于支票结算、现金提取和信用卡交易的机器，再到太空中的地球观测卫星，都在不断收集大量信息。

下面列举一些数据量：

- 目前美国宇航局的地球观测卫星每天生成 1TB 数据。
- 人类基因组计划针对数十亿个遗传碱基中的每一个存储数千个字节。
- 许多公司都维护着大型客户交易数据仓库。一个较小数据仓库可能包含超过一亿个事务。
- 每天在自动记录设备上记录大量数据(如信用卡交易文件和网络日志，以及监控系统记录的非符号数据)。
- 估计有超过 6.5 亿个网站，其中一些网站非常庞大。
- Facebook 拥有超过 9 亿用户，每天估计产生 30 亿个帖子。
- 据估计，Twitter 用户约有 1.5 亿，每天发送 3.5 亿条推文。

随着存储技术的不断发展，无论是商业数据仓库、科研实验室还是其他地方，都越来越多地以较低成本存储大量数据，同时人们逐渐认识到这些数据中可能包含以下知识：对公司的兴衰至关重要的知识，可导致重大科学发现的知识，可更准确地预测

天气和自然灾害的知识，可找出致命疾病原因及治愈方法的知识，以及可能关乎生死的知识。然而，这些数据大部分仅被存储——人们很少对这些数据进行更深入的检查。所以准确地说，世界正变得“数据丰富但知识贫乏”。

机器学习技术有可能解决一直困扰公司、政府和个人的数据爆炸问题。

1.2 知识发现

“知识发现”(Knowledge Discovery)被定义为“从数据中提取隐含的、先前未知的和潜在可用的信息”。该过程虽然只是数据挖掘的一部分，却是核心部分。

图 1.1 显示了完整的知识发现过程的略微理想化的版本。

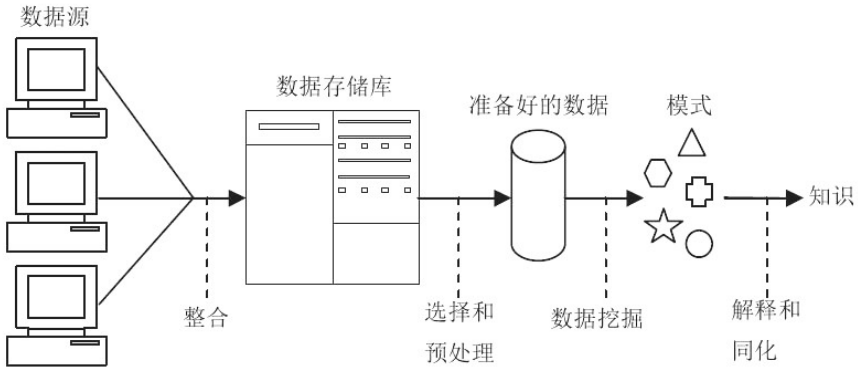


图 1.1　知识发现过程

数据可能有许多来源，被集成并放在一些通用数据存储中。然后将其中一部分数据预处理成标准格式，再将这些“准备好的数据”传递给数据挖掘算法，该算法根据规则或某种其他类型的“模式”生产输出。最后对这些输出进行解释并给出潜在有用的新知识——这就是知识发现的“圣杯”。

通过上面的简述可清楚地看到，数据挖掘算法是知识发现的核心，但并非全部。数据的预处理和结果的解释也非常重要。与其说完成这些任务是一门精确的科学，还不如说是一门艺术。虽然本书也会介绍数据的预处理并解释结果，但重点讨论的是知识发现的数据挖掘阶段涉及的算法。

1.3　数据挖掘的应用

数据挖掘的应用范围越来越广，涉及多个领域：

- 分析卫星图像
- 分析有机化合物
- 自动文摘
- 信用卡欺诈检测
- 电力负荷预测
- 财务预测
- 医疗诊断
- 预测电视观众的比例
- 产品设计
- 房地产估价
- 针对性营销
- 信息检索和文本总结
- 火力发电厂优化
- 毒性危害分析
- 天气预报

此外，还有其他许多应用。潜在或实际的应用示例包括：

- 超市连锁店挖掘客户交易数据，以更快找到高价值客户。
- 信用卡公司可使用客户交易数据仓库进行诈骗检测。
- 大型连锁酒店可使用调查数据库来识别“高价值”客户的特性。
- 通过提高预测不良贷款的能力来预测消费者贷款申请的违约概率。
- 减少 VLSI 芯片的制造缺陷。
- 筛选在半导体制造过程中收集的大量数据，以识别导致问题的条件。
- 预测电视节目的收视率，从而允许管理人员合理安排节目时间，尽量提高收视率并增加广告收入。
- 预测癌症患者对化疗的反应，从而降低医疗费用而不影响护理质量。
- 分析老年人的“动作捕捉”数据。
- 社交网络中的趋势挖掘和可视化。

应用可分为四个主要类型：分类、数值预测、关联规则和聚类。下面简要解释每个类型。但首先需要区分标签和无标签数据。

1.4 标签和无标签数据

一般情况下，会有一个示例数据集(被称为“实例”)，每个实例都包含许多变量的值，这些变量在数据挖掘中通常称为“属性”。共有两类数据，它们以完全不同的方式处理。

对于第一种类型，通常存在一个专门指定的属性，其目的是使用给定数据为未见实例预测该属性的值。这类数据称为“标签数据”(labeled data)。使用标签数据的数据挖掘称为“监督学习”(supervised learning)。如果指定的属性是分类的，即必须取多个不同值中的一个，例如“极好”“好”或“差”，或物体识别应用中的“汽车”“自行车”“人”“公共汽车”或“出租车”，那么该任务被称为“分类”(classification)。如果指定的属性是数字的，例如房屋的预期售价或下一个交易日股票市场的开盘价，那么该任务被称为“回归”(regression)。

没有任何特殊属性的数据称为“无标签数据”，而无标签数据的数据挖掘称为“无监督学习”(unsupervised learning)，目的只是从可用数据中提取尽可能多的信息。

1.5 监督学习：分类

分类是数据挖掘最常见的应用之一，对应于日常生活中经常发生的任务。例如，医院可能希望将患者患某种疾病的风险分类为高、中或低三个类别，民意调查公司可能希望将受访者分类为可能投票给某个政党的人或尚未决定的人，或者希望将学生成绩分类为优秀、良好、合格或不合格。

图 1.2 的示例显示了一种典型情况。该例使用一个表格形式的数据集，其中包含五个科目的学生成绩(属性 SoftEng、ARIN、HCI、CSA 和 Project 的值)及其整体学位类别(Class)。为简单起见，使用点行表示多行。接下来希望找到一些方法，以便根据学生成绩档案来预测学生的整体学位类别。

SoftEng	ARIN	HCI	CSA	Project	Class
A	B	A	B	B	Second
A	B	B	B	B	Second
B	A	A	B	A	Second
A	A	A	A	B	First
A	A	B	B	A	First
B	A	A	B	B	Second
...	...	...	...	...	...
A	A	B	A	B	First

图 1.2　分类数据

可通过多种方法实现上述目标，主要包括：

最近邻匹配。该方法依赖于识别出某种意义上与某个未分类示例“最接近”的五个示例。如果五个“最近邻”具有等级 Second、First、Second、Second 和 Second，就可以合理地得出结论，新实例应该被归类为 Second。

分类规则。寻找可用于预测未见实例分类的相关规则，例如：

IF SoftEng = A AND Project = A THEN Class = First

IF SoftEng = A AND Project = B AND ARIN = B THEN Class = Second

IF SoftEng = B THEN Class = Second

分类树。生成分类规则的一种方法是使用称为“分类树”或“决策树”的中间树状结构。

图 1.3 显示了与学位分类数据对应的决策树。

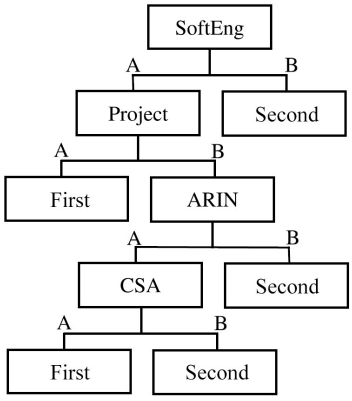


图 1.3　学位分类数据的决策树

1.6　监督学习：数值预测

分类是预测的一种形式，要预测的值是一个标签；而数值预测(通常称为“回归”)是另一种预测形式。此时希望预测一个数值，例如公司的利润或股价。

一种实现数值预测的流行方法是使用神经网络，如图 1.4 所示。

这是一种基于人类神经元模型的复杂建模技术。通常向神经网络提供一组输入并用于预测一个或多个输出。

虽然神经网络是一种重要的数据挖掘技术，但该技术非常复杂，需要用一整本书来介绍，因此在此不再进一步讨论。附录 C 中列出几本介绍神经网络的优秀教材。

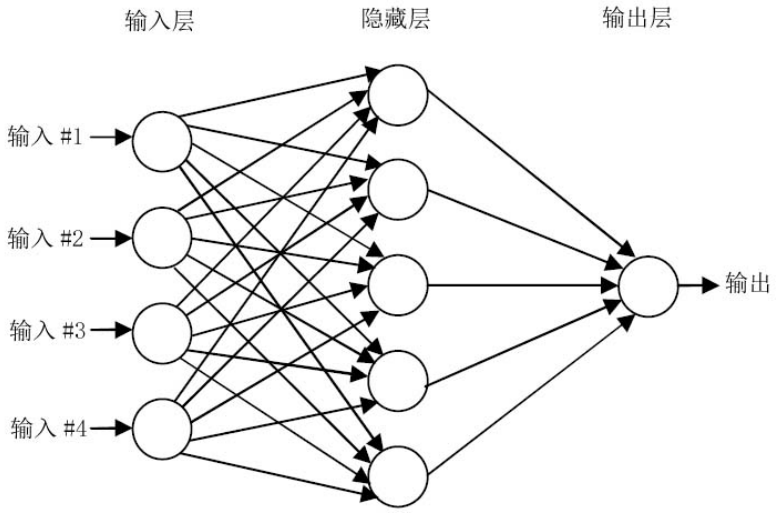


图 1.4　神经网络

1.7　无监督学习：关联规则

有时可能希望使用一个训练集来查找变量值之间存在的任何关系(通常以“关联规则”形式存在)。可从任何给定的数据集中导出许多可能的关联规则。通常会使用一些附加信息来说明关联规则，这些信息表明了关联规则的可靠性，例如：

IF variable _1 > 85 and switch_6 = open
THEN variable_23 < 47.5 and switch_8 = closed(probability = 0.8)

这种应用类型的一种常见形式称为“市场购物篮分析”(market basket analysis)。如果知道所有顾客在一段时间内(如一周内)购买的商品，就可以找到有助于商店在未来更有效地推销其产品的关系。例如：

IF cheese AND milk THEN bread (probability = 0.7)

上述关联规则表明购买奶酪和牛奶的顾客中有 70%也会购买面包，所以为方便顾客，让面包靠近奶酪和牛奶柜台是非常明智的做法；但如果利润更重要，那么明智的做法是将它们分开，以鼓励顾客购买其他商品。

1.8　无监督学习：聚类

聚类算法检查数据以查找相似的项目集。例如，保险公司可根据收入、年龄、购买的保单类型或先前的索赔经验对客户进行分组。而在故障诊断应用中，可根据某些关键变量的值对电机故障进行分组(如图 1.5 所示)。

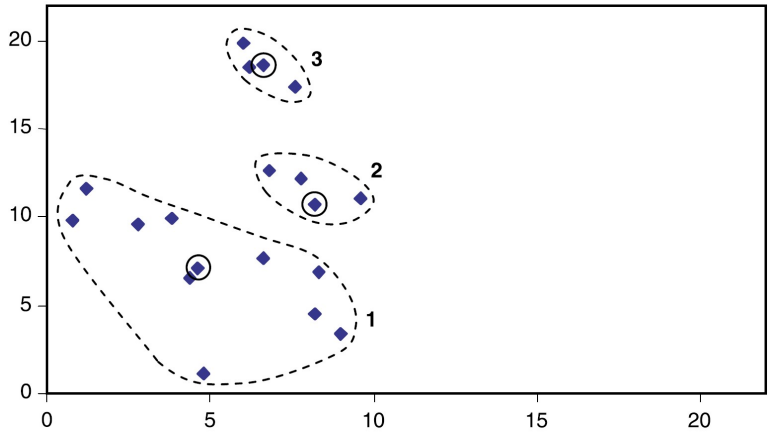


图 1.5　数据聚类

第 2 章

用于挖掘的数据

用于挖掘的数据有多种形式：操作员手动输入的计算机文件、SQL 中的商业信息或其他一些标准数据库格式的信息、由故障记录设备等自动记录的信息以及从卫星发送的二进制数据流。为便于数据挖掘以及帮助读者了解书中介绍的其他内容，假设数据采用 2.1 节中介绍的特定标准形式。

2.1 标准制定

假设对于任何数据挖掘应用程序，都有一系列感兴趣的对象。这通常指一群人(可能是所有活着或逝世的人，又或是医院的所有病人)，也可能是英格兰的所有狗，或无生命的物体(例如从伦敦到伯明翰的所有火车旅行、月球上的所有岩石或万维网上存储的所有页面)。

对象的范围通常非常大，但我们通常只有很小一部分。通常希望从可用的数据中提取信息，并希望这些信息适用于尚未看到的大量数据。

每个对象由与其属性对应的许多“变量”描述。在数据挖掘中，变量通常被称为“属性”。本书将交替使用“变量”和“属性”这两个术语。

对应于每个对象的变量的值集被称为“记录”或“实例”。为应用程序提供的完整数据组被称为“数据集”。数据集通常被描述为一个表，每行代表一个实例，而每列包含每个实例的一个变量(属性)的值。数据集的典型示例是第 1 章中给出的 degrees 数据(如图 2.1 所示)。