计算机科学典范教材

数据挖掘原理

(第 4 版)

[英] 麦克斯 • 布拉默(Max Bramer) 著

李晓峰 逄金辉 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：01-2021-6095

图书在版编目(CIP)数据

数据挖掘原理：第4版 / (英) 麦克斯 • 布拉默(Max Bramer) 著；李晓峰，逄金辉译. —北京：清华大学出版社，2021.12
书名原文：Principles of Data Mining, Fourth Edition
计算机科学典范教材
ISBN 978-7-302-59649-3

Ⅰ. ①数…　Ⅱ. ①麦… ②李… ③逄…　Ⅲ. ①数据采集—教材　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2021)第 257145 号

责任编辑：王　军
装帧设计：孔祥峰
责任校对：成风进
责任印制：沈　露

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：天津鑫丰华印务有限公司
经　　销：全国新华书店
开　　本：170mm×240mm　　**印　　张**：29.75　　**字　　数**：570 千字
版　　次：2022 年 1 月第 1 版　　**印　　次**：2022 年 1 月第 1 次印刷
定　　价：118.00 元

产品编号：089795-01

译 者 序

数据挖掘，又称为资料探勘、数据采矿。它是数据库知识发现(Knowledge-Discovery in Databases，KDD)中的一个步骤。数据挖掘一般指从大量的数据中自动搜索隐藏于其中的有着特殊关系(属于 Association rule learning)的信息的过程。数据挖掘通常与计算机科学有关，并通过统计、在线分析处理、情报检索、机器学习、专家系统(依靠过去的经验法则)和模式识别等诸多方法来实现上述目标。

知识发现过程由以下三个阶段组成：①数据准备，是从相关的数据源中选取所需的数据并整合成用于数据挖掘的数据集；②数据挖掘，是用某种方法将数据集所含的规律找出来；③结果表达和解释，是尽可能以用户可理解的方式(如可视化)将找出的规律表示出来。数据挖掘的任务有关联分析、聚类分析、分类分析、异常分析、特异群组分析和演变分析等。

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。一旦掌握了基本技术，就可通过多种渠道了解该领域的最新进展。本书共 23 章，分别介绍了概述、用于挖掘的数据、朴素贝叶斯和最近邻算法、使用决策树进行分类、决策树归纳、估计分类器的预测精度、连续属性、避免决策树的过度拟合、关于熵的更多信息、归纳分类的模块化规则、度量分类器的性能、处理大量数据、集成分类、比较分类器、关联规则挖掘、聚类、文本挖掘、分类流数据、神经网络。

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。所以本书末尾还有 5 个附录，分别介绍了基本数学知识、数据集、更多信息来源、词汇表和符号、自我评估练习题答案。

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，本书也是极佳的自学书籍。

在这里要感谢清华大学出版社的编辑，他们为本书的翻译投入了巨大的热情并付出了很多心血。没有他们的帮助和鼓励，本书不可能顺利付梓。

对于这本经典之作，译者在翻译的过程中虽力求“信、达、雅”，但由于水平有限，失误在所难免，如有任何意见和建议，请不吝指正。

译　者

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，在学习本书的过程中你不需要拥有太多的数学知识即可理解其中的相关内容。

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。作者以前是一位数学家，他现在喜欢在任何可能的情况下用简单的英语交流，并相信好例子胜过一百个数学符号。

本书涉及数学公式较少，将重点介绍相关概念。但是，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。

封底二维码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，读者可扫描封底二

维码中的参考文献，查阅相关信息。

第 4 版的注意事项

自第 1 版以来，可用于数据挖掘的数据量大幅增加。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。今天，世界上可用的医疗保健数据量估计超过 2 万亿兆字节。为了反映“深度学习”的日益普及，本书新增了最后一章，其中详细介绍了最重要的神经网络类型之一，并展示了如何将其应用于分类任务。

致谢

首先感谢我的女儿 Bryony，她帮助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，她对本书初稿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。

UTiCS

“计算机科学本科生主题”(UTiCS)为计算机和信息科学所有领域的本科生提供高质量的教学内容。UTiCS 书籍采取了新颖、简洁和现代方法，囊括从核心基础和理论材料到最后一年的主题和应用，是自学或一两个学期课程的理想选择。本书由该领域内的知名专家撰写，并由国际顾问委员会审查，包含许多例子和问题，其中许多包括完全有效的解决方案。

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，都在不断地从社交媒体和互联网上收集大量信息。

下面列举一些数据量(当你读到这篇文章时，有些数据已经大大增加了)。

- 目前美国宇航局的地球观测卫星每天生成 1TB 数据。这比以前所有观测卫星所传送的数据总量还要多。
- 生物学家每年产生大约 1 500GB 的基因序列数据。
- 许多公司都维护着大型客户交易数据仓库。一个较小的数据仓库可能包含超过 1 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 CCTV 记录等的非符号数据)。
- 估计有超过 15 亿个网站，其中一些网站非常庞大。
- Facebook 拥有超过 24 亿用户，每天估计上传 3.5 亿张照片。

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探究。所以准确地说，世界正变得“数据丰富但知识贫乏”。

和所有存储的数据一样，每天超过 100 万条记录的数据流(可能永远持续下去)现在也很常见。

机器学习技术(其中一些已经建立了很长时间)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、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理想化版本。

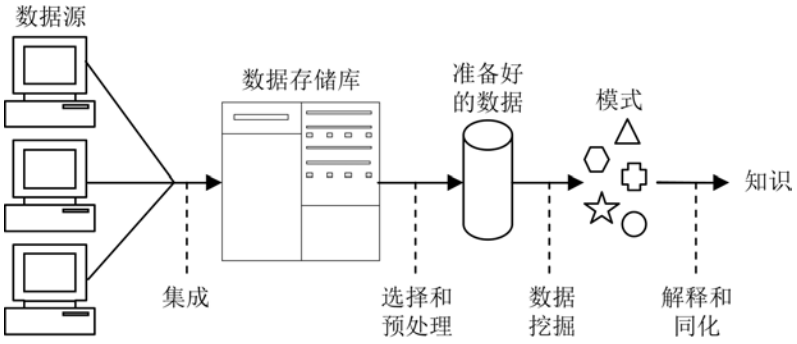

图 1.1 知识发现过程

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，如下所示。

- 分析卫星图像
- 有机化合物分析
- 自动文摘
- 生物信息学
- 信用卡欺诈检测
- 刑事调查
- 客户关系管理
- 电力负荷预测
- 财务预测
- 欺诈检测
- 医疗保健
- 购物篮分析
- 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
- 文本综述
- 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。潜在的或实际的应用示例包括：

- 超市连锁店挖掘客户交易数据，以更快地找到高价值客户。
- 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识别“高价值”客户的特性。
- 通过提高预测不良贷款的能力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。
- 通过分析人脸识别系统的数据，在人群中定位犯罪嫌疑人。
- 分析一系列药物和天然化合物的信息，以确定新抗生素的重要候选药物。
- 分析 MRI 图像以识别可能的脑肿瘤。

应用可分为 4 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。但首先需要区分两种类型的数据。

1.4 标签数据和无标签数据

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被称为“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”。使用标签数据的数据挖掘称为“监督学习”。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”“差”，或物体识别应用中的“汽车”“自行车”“人”“公共汽车”“出租车”，那么该任务被称为“分类”。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”。

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”，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、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、合格和不合格。

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 5 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　学位分类数据

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 5 个示例。如果 5 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。

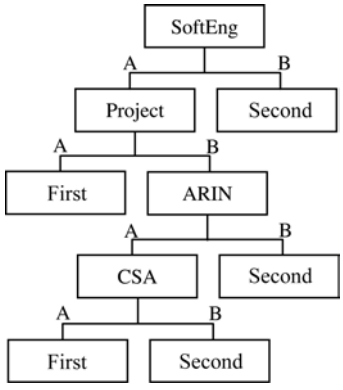

图 1.3　学位分类数据的决策树

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。

第 23 章将讨论最广泛使用的神经网络类型之一。然而，重点主要是分类而不是数值预测。

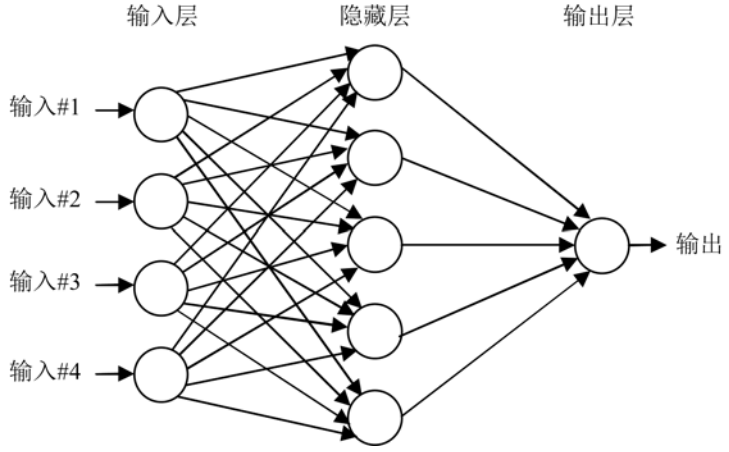

图 1.4　神经网络

1.7 无监督学习：关联规则

有时可能希望使用一个训练集查找变量值之间存在的任何关系(通常以“关联规则”的形式存在)。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说明关联规则，这些信息表明了关联规则的可靠性，如下例：

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”。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，将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(见图 1.5)。

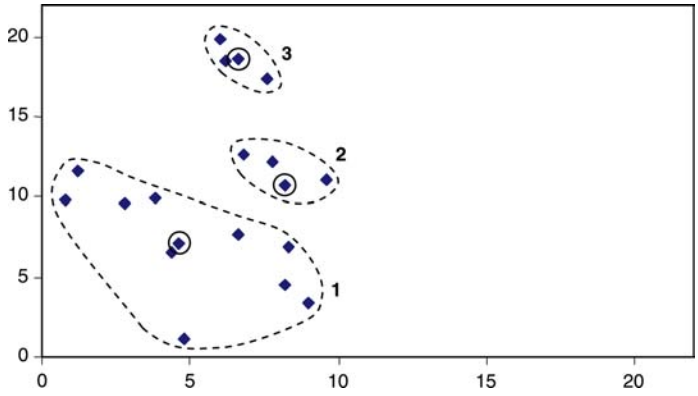

图 1.5　数据聚类