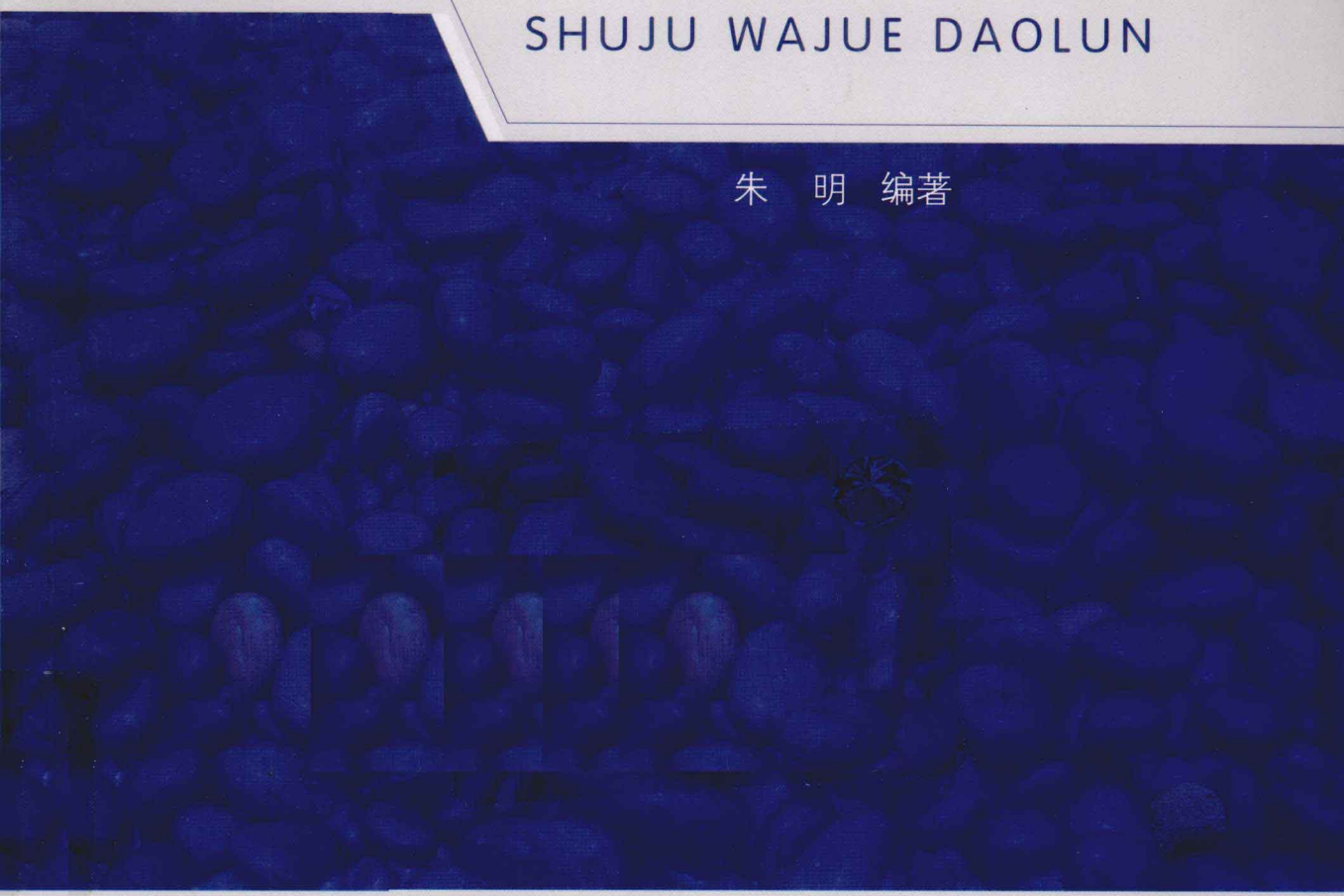
数据挖掘导论

SHUJU WAJUE DAOLUN

朱　明　编著



中国科学技术大学出版社

数据挖掘导论

朱　明　编著

中国科学技术大学出版社

内 容 简 介

与十年前相比，数据挖掘作为数据分析与决策支持的重要技术，已在各行各业得到了更为广泛的应用。随着网络和IT技术的不断发展，数据挖掘应用必将更加深入和普及。作者根据自己十多年教授“数据挖掘”课程的经验积累，编写了这本教材。

本书全面系统地介绍了数据挖掘的主要方法，并配有许多应用案例，使得读者能够更加容易地理解这些数据挖掘方法。同时本书每章后还配有许多思考题，使得这本书更适合作为“数据挖掘”课程的教材。

本书的主要内容包括数据挖掘概述、数据仓库与在线分析、分类挖掘、关联挖掘、聚类挖掘、异类挖掘、数据流挖掘、文本挖掘以及数据挖掘应用与数据挖掘云等。

本书适合作为高等院校高年级本科生、研究生相关课程的教材或参考书。对从事数据挖掘应用的技术人员以及希望了解数据挖掘方法与应用的广大数据挖掘用户，本书也具有一定的参考价值。

图书在版编目(CIP)数据

数据挖掘导论/朱明编著. —合肥：中国科学技术大学出版社，2012.1
ISBN 978-7-312-02958-5

Ⅰ.数… Ⅱ.朱… Ⅲ.数据采集 Ⅳ.TP274

中国版本图书馆CIP数据核字（2011）第265819号

出版 中国科学技术大学出版社
地址：安徽省合肥市金寨路96号，230026
网址：http://press.ustc.edu.cn
印刷 安徽省瑞隆印务有限公司
发行 中国科学技术大学出版社
经销 全国新华书店
开本 787 mm×1092 mm 1/16
印张 19.25
字数 489千
版次 2012年1月第1版
印次 2012年1月第1次印刷
定价 36.00元

前 言

数据挖掘(Data Mining)就是从大量的、有噪声的、不完全的、模糊的甚至随机的实际应用数据中，提取出隐含的、人们事先不知道的但又是潜在有用的知识和信息的过程。

数据挖掘技术从低层的查询到高层的决策支持，应用面十分广泛，吸引了各个领域的人才去研究，从而形成了一门综合了人工智能、数据库技术、统计学、可视化技术、分布式计算等多门学科的交叉学科。

数据挖掘的前身是数据分析，已经有很多年的历史。随着时代的变化，大量数据被收集起来，不仅仅用于分析，更应用于商业运作。分析这些数据已经不仅仅是科学研究的需要，而渐渐在商业中应用起来。数据挖掘技术目前已经大量应用于商业。由于商业的业务量非常大，面对令人眼花缭乱的业务数据，人们往往无法直观上获得有用的信息。而使用数据挖掘技术，对大量的业务数据进行抽取、转换、分析并建立模型，则可从中挖掘出有用的信息，为商业决策提供强有力的支持，大大有利于提高商业竞争力。

数据挖掘作为一门新兴的学科，已表现出强大的应用价值和前景。广大从事数据库应用、机器学习与模式识别、智能计算等诸多领域研究的科技人员迫切需要了解和掌握数据挖掘的方法和技术，为此，作者结合自己十多年“数据挖掘”课程的教学经验，在 2002 年出版、2008 年再版的《数据挖掘》一书基础上，针对“数据挖掘”课程教学和技术人员对数据挖掘基本方法掌握的最新需求，编写完成此书。

本书共分 11 章，主要介绍数据挖掘主要方法、技术与应用等方面的内容。前 6 章介绍主要数据挖掘方法，第 7 章至第 9 章介绍数据挖掘 3 种经典场景下的应用技术，最后两章分别介绍数据挖掘应用案例以及数据挖掘最新发展趋势之一——数据挖掘的云计算实现。

第 1 章主要介绍数据挖掘起源、数据挖掘过程、数据挖掘任务、数据挖掘系统与工具以及数据挖掘发展趋势等内容。

第 2 章主要介绍数据仓库的概念、数据仓库数据模型、数据仓库的构建、数据仓库在线分析以及数据仓库应用示例等内容。

第 3 章主要介绍分类挖掘概述、决策树分类方法、决策树分类算法深入、分类挖掘评估与改进以及分类挖掘的应用等内容。

第 4 章主要介绍贝叶斯分类方法、k 近邻分类方法、神经网络分类方法、遗传算法分类方法、分类器集成方法以及分类挖掘的应用等内容。

第 5 章主要介绍关联挖掘概述、基本关联挖掘方法、关联挖掘深入、分布式关联挖掘和关联挖掘的应用等内容。

第 6 章主要介绍聚类分析的概念、聚类分析中的数据类型、主要聚类方法、基于划分聚类方法、基于层次聚类方法、基于密度聚类方法、基于网格聚类方法、基于模型聚类方法以及聚类挖掘的应用等内容。

第 7 章主要介绍异类挖掘概述、孤立点挖掘方法、基于聚类的异类挖掘、基于数据延续

性的异常挖掘和异类挖掘的应用等内容。

第8章主要介绍数据流挖掘概述、数据流分类挖掘、数据流关联挖掘、数据流聚类挖掘和数据流挖掘的应用等内容。

第9章主要介绍文本挖掘概述、文本表示方法、文本分类挖掘、文本聚类挖掘和文本挖掘的应用等内容。

第10章主要介绍客户关系管理应用、电子商务应用、商务智能应用等内容。

第11章主要介绍云计算概述、分类挖掘云计算、关联挖掘云计算和数据挖掘云的应用等内容。

尽管作者付出了诸多努力，但由于本人掌握的知识及水平有限，加之数据挖掘技术发展迅速，涉及内容较为广泛，书中不足和错误之处在所难免，恳请广大读者和同行专家批评指正。

朱 明

2011年5月10日于合肥

目　录

前言 …… (Ⅰ)

第 1 章　数据挖掘导论 …… (1)
1.1　数据挖掘的起源 …… (1)
1.2　数据挖掘的过程 …… (4)
1.3　数据挖掘的任务 …… (6)
1.4　数据挖掘系统与工具 …… (10)
1.5　数据挖掘的发展趋势 …… (12)
本章小结 …… (15)
思考题 …… (15)

第 2 章　数据仓库与在线分析 …… (16)
2.1　数据仓库的概念 …… (16)
2.2　数据仓库数据模型 …… (20)
2.3　数据仓库的构建 …… (24)
2.4　数据仓库在线分析 …… (29)
2.5　数据仓库应用示例 …… (35)
本章小结 …… (42)
思考题 …… (43)

第 3 章　分类挖掘(1) …… (44)
3.1　分类挖掘概述 …… (44)
3.2　决策树分类方法 …… (46)
3.3　决策树分类算法深入 …… (52)
3.4　分类挖掘评估与改进 …… (56)
3.5　分类挖掘应用 …… (62)
本章小结 …… (68)
思考题 …… (69)

第 4 章　分类挖掘(2) …… (70)
4.1　贝叶斯分类方法 …… (70)
4.2　k 近邻分类方法 …… (76)
4.3　神经网络分类方法 …… (79)
4.4　遗传算法分类方法 …… (84)
4.5　分类器集成方法 …… (92)

4.6 分类挖掘应用 …… (97)
本章小结 …… (100)
思考题 …… (101)

第5章 关联挖掘 …… (102)
5.1 关联挖掘概述 …… (102)
5.2 基本关联挖掘方法 …… (104)
5.3 关联挖掘深入 …… (114)
5.4 分布式关联挖掘 …… (117)
5.5 关联挖掘应用 …… (121)
本章小结 …… (127)
思考题 …… (128)

第6章 聚类分析 …… (130)
6.1 聚类分析概述 …… (130)
6.2 聚类分析中的数据类型 …… (132)
6.3 主要聚类方法 …… (138)
6.4 划分方法 …… (140)
6.5 层次方法 …… (144)
6.6 基于密度方法 …… (150)
6.7 基于网格方法 …… (153)
6.8 基于模型方法 …… (155)
6.9 聚类挖掘应用 …… (159)
本章小结 …… (163)
思考题 …… (164)

第7章 异类挖掘 …… (165)
7.1 异类挖掘概述 …… (165)
7.2 孤立点挖掘方法 …… (169)
7.3 基于聚类的异类挖掘 …… (174)
7.4 基于数据延续性的异常挖掘 …… (178)
7.5 异类挖掘应用 …… (181)
本章小结 …… (186)
思考题 …… (186)

第8章 数据流挖掘 …… (188)
8.1 数据流挖掘概述 …… (188)
8.2 数据流分类挖掘 …… (193)
8.3 数据流关联挖掘 …… (196)
8.4 数据流聚类挖掘 …… (204)
8.5 数据流挖掘应用 …… (208)
本章小结 …… (213)

思考题 …… (214)

第 9 章　文本挖掘 …… (215)

9.1　文本挖掘概述 …… (215)

9.2　文本表示方法 …… (219)

9.3　文本分类挖掘 …… (225)

9.4　文本聚类挖掘 …… (231)

9.5　文本挖掘应用 …… (234)

本章小结 …… (245)

思考题 …… (246)

第 10 章　数据挖掘应用 …… (247)

10.1　客户关系管理应用 …… (247)

10.2　电子商务应用 …… (256)

10.3　商务智能应用 …… (264)

本章小结 …… (271)

思考题 …… (272)

第 11 章　数据挖掘云 …… (273)

11.1　云计算概述 …… (273)

11.2　分类挖掘云计算 …… (279)

11.3　关联挖掘云计算 …… (284)

11.4　数据挖掘云应用 …… (288)

本章小结 …… (296)

思考题 …… (297)

参考文献 …… (298)

第 1 章　数据挖掘导论

随着计算机软硬件技术的发展，尤其是计算机网络的发展与普及，计算机处理和存储的数据，正在以难以预计的速度增长；另一方面，随着社会经济的不断发展，商业竞争日趋白热化，人们迫切需要从数据中获得有用的知识来帮助进行科学决策。针对“数据丰富而知识贫乏”这一窘境，数据挖掘应运而生。

1.1　数据挖掘的起源

近十几年来，人们利用信息技术生产和收集数据的能力大幅度提高，无数个数据库被用于商业管理、政府办公、科学研究和工程开发等，这一势头仍将持续发展下去。于是，一个新的挑战被提了出来：在这个被称之为“信息爆炸”的时代，信息过量几乎成为人人需要面对的问题。如何才能不被信息的汪洋大海所淹没，从中及时发现有用的知识，提高信息利用率呢？要想使数据真正成为一个公司的资源，只有充分利用它为公司自身的业务决策和战略发展服务才行，否则大量的数据反而可能成为包袱，甚至成为垃圾。因此，面对“数据丰富而知识贫乏”的挑战，如图 1.1 所示，数据挖掘技术应运而生，并得以蓬勃发展，越来越显示出其强大的生命力。

图 1.1　数据丰富但知识缺乏

数据挖掘(Data Mining,DM),又称数据库中的知识发现(Knowledge Discover in Database,KDD),就是从大量的、不完全的、有噪声的、模糊的甚至随机的实际应用数据中,提取出隐含在其中的、人们事先不知道的但又是潜在有用的信息和知识的过程。

数据挖掘是目前人工智能和数据库领域研究的热点问题,它是一种决策支持过程,主要基于人工智能、机器学习、模式识别、统计学、数据库、可视化等技术,自动化地分析企业的数据,做出归纳性的推理,从中挖掘出潜在的模式,为决策者调整市场策略,减少风险,做出正确的决策提供知识支持。

从广义上理解,数据、信息也是知识的表现形式,但是人们更愿意把概念、规则、模式、规律和约束等看作是知识。人们把数据看作是形成知识的源泉,好像从矿石中采矿或淘金一样。原始数据可以是结构化的,如关系数据库中的数据;也可以是半结构化的,如文本、图形和图像数据;甚至是分布在网络上的异构型数据。

数据挖掘是一种新的商业信息处理技术,其主要特点是对商业数据库中的大量业务数据进行抽取、转换、分析和其他模型化处理,从中提取出辅助商业决策的关键性数据。

数据挖掘其实就是一类深层次的数据分析方法。数据分析本身已经有很多年的历史,只不过在过去数据收集和分析的目的是用于科学研究,另外,由于当时计算能力的限制,对大数据量进行分析的复杂数据分析方法受到很大限制。现在,由于各行业业务自动化的实现,商业领域产生了大量的业务数据,这些数据不再是为了分析的目的而收集的,而是由于纯商业运作而产生。分析这些数据也不再是单纯为了研究的需要,更主要的是为商业决策提供真正有价值的信息,进而获得利润。但所有企业面临的一个共同问题是:企业数据量非常大,而其中真正有价值的信息却很少,因此从大量的数据中经过深层分析,获得有利于商业运作、提高竞争力的信息,就像从矿石中淘金一样。

因此,数据挖掘又可以理解为是一种按企业既定业务目标,对大量的企业数据进行探索和分析,以揭示隐藏的、未知的或验证已知的规律性,并进一步将其模型化的先进有效的方法。

随着数据挖掘研究逐步走向深入,人们越来越清楚地认识到,数据挖掘的研究主要涉及数据库、人工智能和数理统计三个领域。数据库技术在经过了 20 世纪 80 年代的辉煌之后,已经在各行各业成为一种文化或时尚,数据库界目前除了关注分布式数据库、面向对象数据库、多媒体数据库、查询优化和并行计算等技术外,已经在开始反思:数据库实质的应用仅仅是查询吗? 理论根基最深的关系型数据库最本质的技术进步点,就是数据存放和数据使用之间的相互分离。查询是数据库的奴隶,发现才是数据库的主人;数据只为普通职员服务,不为管理者决策服务!

由于数据库文化的迅速普及,用数据库作为知识源具有坚实的基础;另一方面,对于一个感兴趣的特定领域——客观世界,先用数据库技术将其形式化并组织起来,就会大大提高知识获取起点,以后从中发掘或发现的所有知识都是针对该数据库而言的。因此,在需求的驱动下,很多数据库学者转向了对数据仓库和数据挖掘的研究,从对演绎数据库的研究转向对归纳数据库的研究。

人工智能学者开始着手基于案例的推理,尤其是从事机器学习的科学家们,不再满足自己构造的小样本学习模式的象牙塔,开始正视现实生活中大量的、不完全的、有噪声的、模糊的、随机的大数据样本,也走上了数据挖掘的道路。

数理统计是应用数学中最重要、最活跃的学科之一,它在计算机发明之前就诞生了,迄

今已有几百年的发展历史。如今相当强大而有效的数理统计方法和工具，已成为信息咨询业的基础。信息时代，咨询业更为发达。然而，数理统计和数据库技术结合得并不算紧密，数据库查询语言 SQL 中的聚合函数功能极其简单，就是一个证明。咨询业用数据库查询数据还远远不够。一旦人们有了从数据查询到知识发现、从数据演绎到数据归纳的要求，概率论和数理统计就获得了新的生命力，所以才会在数据挖掘这个结合点上，立即呈现出“忽如一夜春风来，千树万树梨花开”的繁荣景象。

在互联网上有不少数据挖掘(DM)电子出版物，其中以半月刊“Knowledge Discovery Nuggets”最为权威，如要免费订阅，只需向 http://www.kdnuggets.com/subscribe.html 发送一份电子邮件即可，还可以下载各种各样的数据挖掘工具软件和典型的样本数据仓库，供测试和评价。

数据挖掘的应用非常广泛，下面介绍的一些应用可以说明：

(1) 商品销售。商业部门把数据视作是一种竞争性的财富，为此需要把大型市场营销数据库演变成一个数据挖掘系统。科拉福特食品公司(KGF)是应用市场营销数据库的公司之一，该公司收集了购买过它商品的 3000 万个用户的名单，这是 KGF 通过各种促销手段得到的。KGF 定期向这些用户发送名牌产品的优惠券，介绍新产品的性能和使用情况。该公司体会到了解自己商品的用户越多，则购买和使用这些商品的机会也就越多，公司的营业状况也就越好。

(2) 旅行服务。美国企业 Travel Wind 一直以来主要的服务内容是提供基础性的旅行服务，诸如机票预定、旅馆预定等。公司现在考虑向部分国内外客户提供更高附加值的服务项目，为了准确了解即将推出的一些高附加值产品和服务真实潜在需求的实际状况，公司对部分客户进行了一次客户问卷调查。通过对回收问卷的数据整理以及随后的数据挖掘工作，一些有价值的市场营销线索脱颖而出。通过聚类挖掘发现存在两个群体——大众消费群体和高端群体，为此 Travel Wind 公司得出以下营销线索总结：第一，本公司现在有足够的机会和机遇开展高端产品和服务的推介，因为超过一半以上的公司客户都是经常在国内或国外旅行的；第二，本公司可以考虑向两个不同的客户群体分别开发两种不同的营销创意广告宣传语(基于他们不同的收入和年龄特点)。高端客户似乎是处于职业顶点的专业人士或者是富裕的家庭主妇，而大众客户似乎是退休人士在寻找独特的旅游经历。

(3) 金融服务/信用卡。通用汽车公司(General Motors)已经采用信用卡——GM 卡，在该公司的数据库中已拥有 1200 万个持有信用卡的客户。公司通过观察，可以了解客户正在驾驶什么样的汽车，下一步计划购买什么样的汽车及他们喜欢哪一类车辆。

(4) 美国的 AutoTrader.com 是世界上最大的汽车销售站点，每天都会有大量的用户对其网站上的信息进行点击，他们运用 SAS 软件进行数据挖掘，每天对数据进行分析，找出用户的访问模式，就用户对产品的喜欢程度进行判断，并设立特定服务，从而取得了成功。

(5) Reuteres 是世界著名的金融信息服务公司，其利用的数据大都是外部的数据，这样数据的质量就是公司生存的关键所在，必须从数据中检测出错误的成分。Reuteres 利用 SPSS 的数据挖掘工具 SPSS/Clementine，建立起数据挖掘模型，极大地提高了错误检测率，保证了信息的正确性和权威性。

1.2 数据挖掘的过程

在数据挖掘中，被研究的业务对象是整个过程的基础，它驱动了整个数据挖掘过程，也是检验最后结果和指引分析人员完成数据挖掘的依据和顾问。图 1.2 中各步骤是按一定顺序完成的，当然整个过程中还会存在步骤间的反馈。数据挖掘的过程并不是自动的，绝大多数的工作需要人工完成。在整个数据挖掘过程中，60%的时间用在数据准备上，这说明了数据挖掘对数据的严格要求，而后续挖掘工作仅占总工作量的 10%。

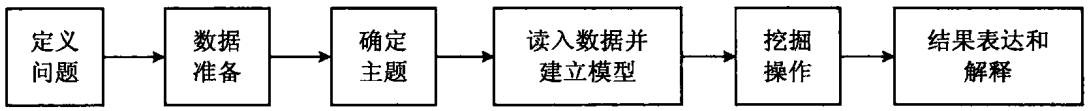


图 1.2 数据挖掘的一般流程

从大量的、不完全的、有噪声的、模糊的甚至随机的实际应用数据中提取出隐含在其中而又非常有用的信息、模式(规则)和趋势的数据挖掘过程主要包括 6 个步骤，各步骤的大体内容如下：

(1) 定义问题。首先明确定义将要解决的问题。数据挖掘者要熟悉所研究行业的数据和业务问题，缺乏这些，就不能够充分发挥数据挖掘的价值，很难得到正确的结果。模型的建立取决于问题的定义，有时相似的问题，所要求的模型几乎完全不同。

清晰地定义出业务问题，认清数据挖掘的目的，是数据挖掘的重要一步。挖掘的最后结果是不可预测的，但要探索的问题应是有预见的，为了数据挖掘而数据挖掘则带有盲目性，是不会成功的。

(2) 数据准备。有些人将数据挖掘看作是一个不可思议的过程，认为它吞进的是原始数据，吐出来的是“钻石”。数据准备正是这个过程的核心。这一阶段又可分为 3 个子步骤：数据集成，数据选择，数据预处理。数据集成将多文件或多数据库运行环境中的数据进行合并处理，解决语义模糊性，处理数据中的遗漏和清洗脏数据等。数据选择的目的是辨别出需要分析的数据集合，缩小处理范围，提高数据挖掘的质量，因此需要搜索所有与业务对象有关的内部和外部数据信息，并从中选择出适用于数据挖掘应用的数据。而数据预处理则是为了克服目前数据挖掘工具的局限性，提高数据质量，同时将数据转换成一个适用于特定挖掘算法的分析模型。建立一个真正适合挖掘算法的分析模型是数据挖掘成功的关键。

(3) 确定主题。数据挖掘过程的第三步是确定研究主题。数据挖掘是一个经常需要回溯的过程，因此没有必要在数据完全准备好之后才开始进行数据挖掘。随着时间的推移，你所使用的数据、你对它们分组的方式以及数据清洗的效果等都将改变，并有可能改进整个模型。这一步会涉及了解研究主题的局限性，选择待完成的良好研究主题，确定待研究的合适的数据元素，以及决定如何进行数据操作等。

(4) 读入数据并建立模型。一旦确定要输入的数据之后，接着就是要用数据挖掘工具读入数据并从中构造出一个模型。根据所选用的数据挖掘工具的不同，所构造出的数据模型也会有很大的差别。

(5) 挖掘操作。依照上述准备工作，利用选好的数据挖掘工具在数据中查找。这个搜索过程可以由系统自动执行，自底向上搜索原始事实以发现它们之间的某种联系，也可以加入用户交互过程，由分析人员主动发问，从上到下地找寻以验证假设的正确性。数据挖掘的搜索过程需要反复多次，通过评价数据挖掘结果不断调整数据挖掘的精度，以达到发现知识的目的。

(6) 结果表达和解释。根据最终用户的决策目标对提取出的信息进行分析，把最有价值的信息区分出来，并通过决策支持工具提交给决策者。

数据挖掘过程的分步实现，不同的阶段会需要有不同专长的人员，他们大体可以分为三类：

(1) 业务分析人员：要求精通业务，能够解释业务对象，并能根据各业务对象确定出用于数据定义和挖掘算法的业务需求。

(2) 数据分析人员：要求精通数据分析技术，对统计学有较熟练的掌握，有能力把业务需求转化为数据挖掘的各步操作，并为每步操作选择合适的技术。

(3) 数据管理人员：要求精通数据管理技术，并能从数据库或数据仓库中搜集数据。

从上可见，数据挖掘是一个多种专业人员相互配合的工作过程，也是一个在资金上和技术上高投入的过程。这一过程要反复进行，在反复的过程中，不断地趋近事物的本质，不断地优选问题的解决方案。

20 世纪 90 年代后期，当时的数据挖掘市场是年轻而不成熟的，但是这个市场显示出了爆炸式的增长。3 个在这方面经验丰富的公司 Daimler Chrysler、SPSS、NCR 发起并建立了一个社团，目的是建立数据挖掘方法和过程的标准。在获得了 EC(European Commission)的资助后，他们开始实现他们的目标。为了征集业界广泛的意见，共享知识，他们创建了 CRISP-DM Special Interest Group(简称为 SIG)。SIG 组织开发并提炼出 CRISP-DM (Cross-Industry Standard Process for Data Mining)，如图 1.3 所示，同时在 Mercedes-Benz 和 OHRA(保险领域企业)中进行了大规模数据挖掘项目的实际试用。SIG 还将 CRISP-DM 和商业数据挖掘工具集成起来。SIG 组织目前在伦敦、纽约、布鲁塞尔已经发展到 200 多个成员。

当前 CRISP-DM 提供了一个数据挖掘生命周期的全面评述，包括项目的相应周期、它们的各自任务和这些任务的关系。在这个描述层中，识别出所有关系是不可能的。所有数据挖掘任务之间关系的存在依赖于用户的目的、背景和兴趣，最重要的还有数据。SIG 组织已经发布了 CRISP-DM Version 1.0 Process Guide and User Manual 的电子版，这个可以免费使用。

一个数据挖掘项目的生命周期包含 6 个阶段。这 6 个阶段的顺序是不固定的，我们经常需要前后调整这些阶段。这依赖于每个阶段或是阶段中特定任务的产出物是否是下一个阶段必需的输入。图 1.3 中的箭头指出了最重要的和依赖度高的阶段关系。

图 1.3 中的外圈象征数据挖掘自身的循环本质——在一个解决方案发布之后一个数据挖掘的过程才可以继续。在这个过程中得到的知识可以触发新的、经常是更聚焦的商业问题。后续的过程可以从前一个过程中得到益处。

(1) 业务理解(Business Understanding)。最初的阶段集中在理解项目目标和从业务的角度理解需求，同时将这个知识转化为数据挖掘问题的定义和完成目标的初步计划。

(2) 数据理解(Data Understanding)。数据理解阶段从初始的数据收集开始，通过一些

活动的处理，以熟悉数据，识别数据的质量问题，首次发现数据的内部属性，或是探究引起兴趣的子集以形成隐含信息的假设。

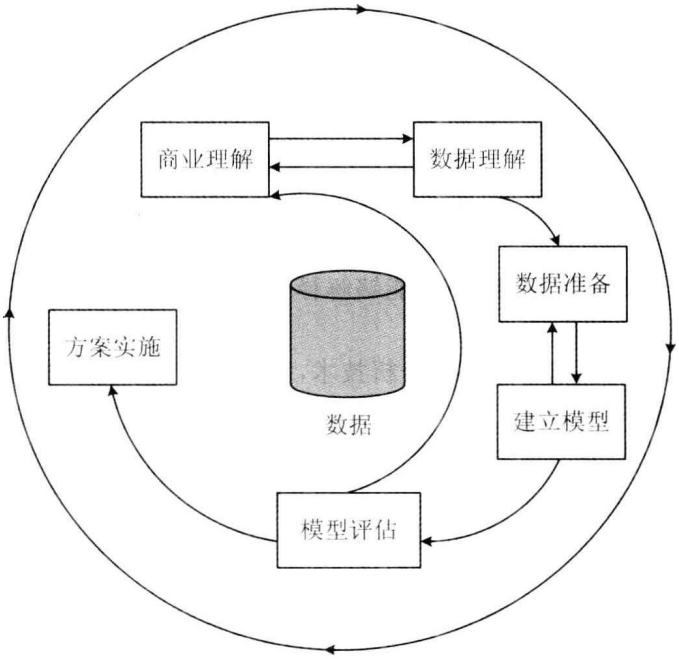


图 1.3 CRISP-DM 的组成架构

(3) 数据准备(Data Preparation)。数据准备阶段包括从未处理数据中构造最终数据集的所有活动。这些数据将是模型工具的输入值。这个阶段的任务有可能执行多次，没有任何规定的顺序。任务包括表、记录和属性的选择，模型工具的转换和数据的清洗。

(4) 建模(Modeling)。在这个阶段，可以选择和应用不同的模型技术，模型参数被调整到最佳的数值。有些技术可以解决一类相同的数据挖掘问题。有些技术在数据形成上有特殊要求，因此需要经常跳回到数据准备阶段。

(5) 评估(Evaluation)。到项目的这个阶段，就已经从数据分析的角度建立了一个高质量显示的模型。在开始最后部署模型之前，重要的事情是彻底地评估模型，检查构造模型的步骤，确保模型可以完成业务目标。这个阶段的关键任务是确定是否有重要业务问题没有被充分地考虑。在这个阶段结束后，必须达成一个数据挖掘结果使用的决定。

(6) 部署(Deployment)。通常，模型的创建不是项目的结束。模型的作用是从数据中找到知识，获得的知识需要以便于用户使用的方式重新组织和展现。根据需求，这个阶段可以产生简单的报告，或实现一个比较复杂的、可重复的数据挖掘过程。在很多案例中，这个阶段是由客户而不是数据分析人员承担部署的工作。

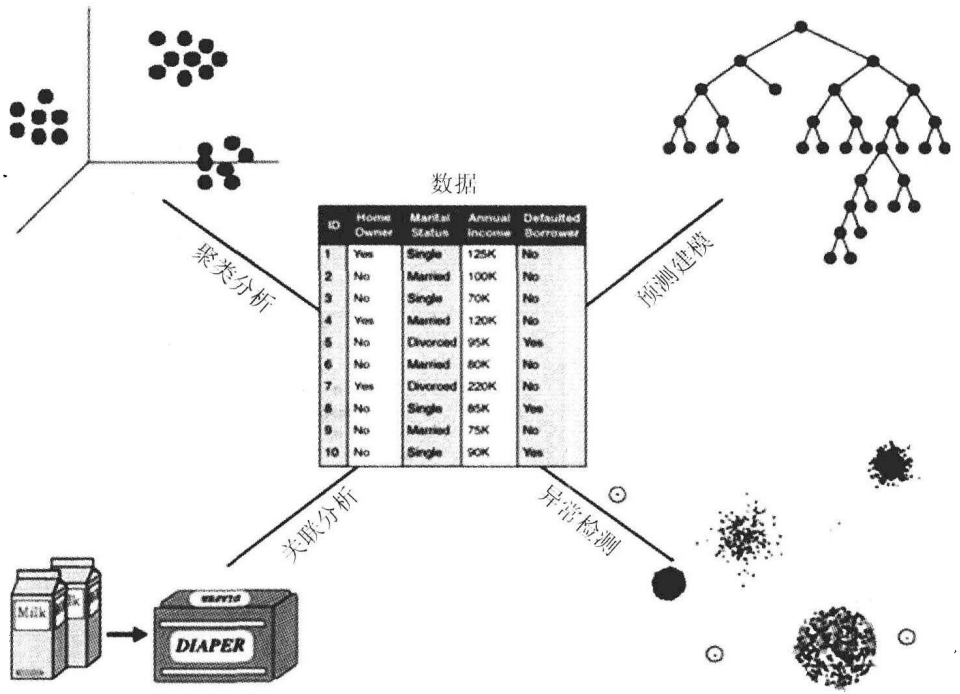
1.3 数据挖掘的任务

通常，数据挖掘的任务可以分为以下两大类：

(1) 预测任务。这些任务的目标是根据其他属性的值，预测特定属性的值。被预测的属性一般称为目标变量(Target Variable)或因变量(Dependent Variable)，而用来做预测的属性称为说明变量(Explanatory Variable)或自变量(Independent Variable)。

(2) 描述任务。其目标是导出概括数据中潜在联系的模式(相关、趋势、聚类、轨迹和异常)。本质上，描述性数据挖掘任务通常是探查性的，并且常常需要后处理技术验证和解释结果。

图 1.4 展示了 4 种主要的数据挖掘任务。



ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

图 1.4　4 种主要的数据挖掘任务

(1) 预测建模(Predictive Modeling)

涉及以说明变量函数的方式为目标变量建立模型。有两类预测建模任务：分类(Classification)，用于预测离散的目标变量；回归(Regression)，用于预测连续的目标变量。例如，预测一个 Web 用户是否会在网上书店买书是分类任务，因为该目标变量是二值的；而预测某股票的未来价格则是回归任务，因为价格具有连续值属性。两项任务的目标都是训练一个模型，使目标变量预测值与实际值之间的误差达到最小。预测建模可以用来确定顾客对产品促销活动的反应，预测地球生态系统的扰动，或根据检查结果判断病人是否患有某种疾病，等等。

例 1.1　预测花的类型。

考虑如下任务：根据花的特征预测花的种类。本例考虑根据是否属于 Setosa、Versicolour、Virginica 这三类之一对鸢尾花(Lris)进行分类。为进行这一任务，我们需要一个数据集，包含这三类花的特性。一个具有这类信息的数据集是著名的鸢尾花数据集，可从加州大学欧文分校的机器学习数据库中得到(http://www.ics.uci.edu/～mlearn)。除花的种类之外，该数据集还包含萼片宽度、萼片长度、花瓣长度、花瓣宽度 4 个其他属性。图 1.5 中给

出了鸢尾花数据集中 150 种花的花瓣宽度与花瓣长度的对比图。花瓣宽度分成 low、medium、high 三类，分别对应于区间[0,0.75)、[0.75,1.75)、[1.75,2.5)。花瓣长度也分成 low、medium、high 三类，分别对应于区间[0,2.5)、[2.5,5)、[5,7)。根据花瓣宽度和长度的这些类别，可以推出如下规则：

- 花瓣宽度和花瓣长度为 low 蕴涵 Setosa。
- 花瓣宽度和花瓣长度为 medium 蕴涵 Versicolour。
- 花瓣宽度和花瓣长度为 high 蕴涵 Virginica。

尽管这些规则不能对所有的花进行分类，但是已经可以对大多数花很好地进行分类(尽管不完善)。注意：根据花瓣宽度和花瓣长度，Setosa 种类的花完全可以与 Versicolour 和 Virginica 种类的花分开，但是后两类花在这些属性上有一些重叠。

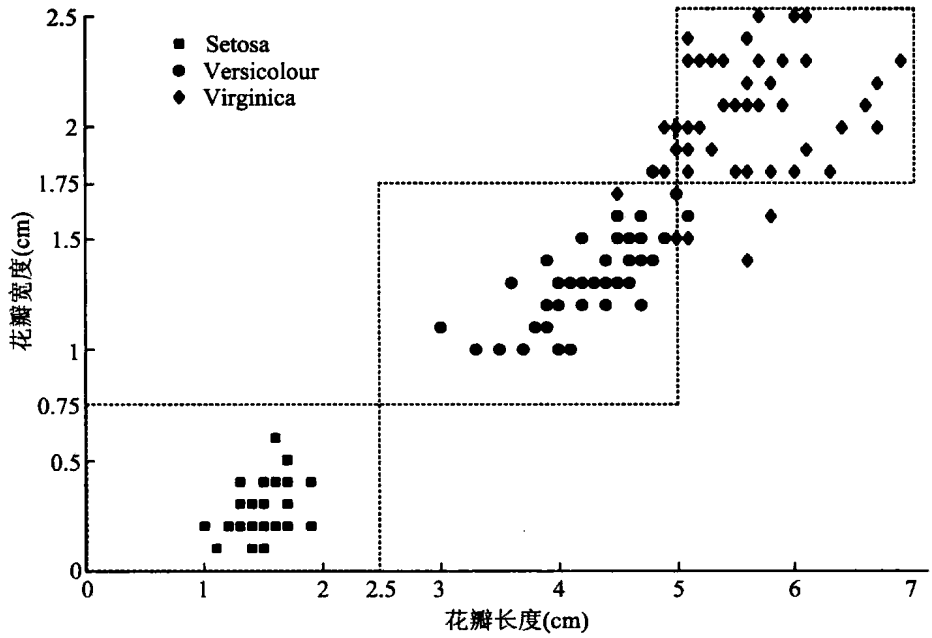


图 1.5 鸢尾花数据集中 150 种花的花瓣宽度与花瓣长度的对比图

(2) 关联分析(Association Analysis)

用来发现描述数据中强关联特征的模式。所发现的模式通常用蕴涵规则或特征子集的形式表示。由于搜索空间是指数规模的，关联分析的目标是以有效的方式提取最有趣的模式。关联分析的应用包括找出具有相关功能的基因组、识别用户一起访问的 Web 页面、理解地球气候系统不同元素之间的联系等。

例 1.2 购物篮分析。

表 1.1 给出的事务是在一家杂货店收银台收集的销售数据。关联分析可以用来发现顾客经常同时购买的商品。例如，我们可能发现规则{尿布}→{牛奶}。该规则暗示购买尿布的顾客多半会购买牛奶。这种类型的规则可以用来发现各类商品中可能存在的交叉销售的商机。

表 1.1　购物篮数据

事务 ID	商　品
1	{面包,黄油,尿布,牛奶}
2	{咖啡,糖,小甜饼,鲑鱼}
3	{面包,黄油,咖啡,尿布,牛奶,鸡蛋}
4	{面包,黄油,鲑鱼,鸡}
5	{鸡蛋,面包,黄油}
6	{鲑鱼,尿布,牛奶}
7	{面包,茶,糖,鸡蛋}
8	{咖啡,糖,鸡,鸡蛋}
9	{面包,尿布,牛奶,盐}
10	{茶,鸡蛋,小甜饼,尿布,牛奶}

(3) 聚类分析(Cluster Analysis)

聚类分析旨在发现紧密相关的观测值组群,使得与属于不同簇的观测值相比,属于同一簇的观测值相互之间尽可能类似。聚类可用来对相关的顾客分组、找出显著影响地球气候的海洋区域以及压缩数据等。

例 1.3　文档聚类。

表 1.2 给出的新闻文章可以根据它们各自的主题分组。每篇文章表示为词—频率对(w,c)的集合,其中 w 表示词,而 c 是该词在文章中出现的次数。在该数据集中,有两个自然簇。第一个簇由前 4 篇文章组成,对应于经济新闻;而第二个簇包含后 4 篇文章,对应于卫生保健新闻。一个好的聚类算法应当能够根据文章中出现的词的相似性,识别这两个簇。

表 1.2　新闻文章集合

文章	词
1	dollar: 1, industry: 4, country: 2, loan: 3, deal: 2, government: 2
2	machinery: 2, labor: 3, market: 4, industry: 2, work: 3, country: 1
3	job: 5, indication: 3, rise: 2, jobless: 2, market: 3, country: 2, index: 3
4	domestic: 3, forecast: 2, gain: 1, market: 2, sale: 3, price: 2
5	patient: 4, symptom: 2, drug: 3, health: 2, clinic: 2, doctor: 2
6	pharmaceutical: 2, company: 3, drug: 2, vaccine: 1
7	death: 2, cancer: 4, drug: 3, public: 4, health: 3, director: 2
8	medical: 2, cost: 3, increase: 2, patient: 2, health: 3, care: 1

(4) 异常检测(Anomaly Detection)

异常检测的任务是识别其特征显著不同于其他数据的观测值。这样的观测值称为异常点(Anomaly)或离群点(Outlier)。异常检测算法的目标是发现真正的异常点,而避免错误地将正常的对象标注为异常点。换言之,一个好的异常检测器必须具有高检测率和低误报率。异常检测的应用包括检测欺诈、网络攻击、疾病的不寻常模式、生态系统扰动等。