数据挖掘与机器学习

徐雪琪　徐蔼婷　编著

清華大學出版社
北　京

内 容 简 介

本书以应用为导向介绍数据挖掘与机器学习相关理论与方法，包括概述、数据与数据平台、数据预处理与特征工程、关联分析、决策树、集成学习、贝叶斯分类、神经网络与深度学习等相关理论及经典算法，以及相关实践案例。本书所有案例均通过 R 或 Python 实现，同时包含详细的分析过程和可视化内容。本书可作为统计学、数据科学与大数据等相关专业高年级本科生和硕士研究生的数据挖掘与机器学习相关课程的教材，也可作为其他数据挖掘与机器学习爱好者的参考用书。

图书在版编目(CIP)数据

数据挖掘与机器学习 / 徐雪琪, 徐蔼婷编著.
北京 : 清华大学出版社, 2025. 7. -- ISBN 978-7-302
-69658-2

Ⅰ. TP311.131；TP181

中国国家版本馆 CIP 数据核字第 20256TK728 号

责任编辑：高 巍
封面设计：马筱琨
版式设计：思创景点
责任校对：马遥遥
责任印制：刘 菲

出版发行：清华大学出版社
网 址：https://www.tup.com.cn，https://www.wqxuetang.com
地 址：北京清华大学学研大厦 A 座 **邮 编：**100084
社 总 机：010-83470000 **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：三河市人民印务有限公司
经 销：全国新华书店
开 本：185mm×260mm **印 张：**17.75 **字 数：**421 千字
版 次：2025 年 8 月第 1 版 **印 次：**2025 年 8 月第 1 次印刷
定 价：69.00 元

产品编号：109902-01

前　言

在数字化浪潮席卷全球的今天，数据已成为驱动社会发展的核心要素。我国在“十四五”规划中明确提出加快数字化发展，推动人工智能、大数据等前沿技术与实体经济深度融合。数据挖掘与机器学习作为这一进程的核心技术，其重要性不言而喻。

本教材是浙江省登峰学科(浙江工商大学统计学)、国家一流本科专业建设点(经济统计学)、浙江省大数据专业教材研究基地、浙江省普通本科高校“十四五”重点立项建设教材的建设成果之一，具有以下显著特点。

(1) 编写风格简洁明了，结构清晰。本教材每章的知识导图将教材中的重要概念和关键内容以图形化方式显示，从而更直观地呈现知识结构和逻辑。同时，本教材注重阐述关键概念和算法的基本思想，避免过度的公式推导，使读者更容易理解和掌握。

(2) 注重实践，涵盖全流程知识。实践的观点是马克思主义哲学的核心观点，本教材注重实践，不仅阐述了数据挖掘和机器学习的经典理论与方法，还涵盖了实践全流程所需的知识，包括数据类型与存储环境、大数据平台(采集、存储、处理与分析)、预处理与特征工程常用的方法等。

(3) 强化育人功能，注重个性化发展。本教材在内容安排上将价值性与知识性相统一，每章以与该章知识紧密相联的导读开篇，引导读者从国家需求、行业痛点和社会价值等维度思考问题。在个性化发展方面，本教材安排了 R 与 Python 两类工具的实践案例，包含详细的分析过程和可视化内容；每章末尾的“拓展”部分，提出了可进一步学习的不同方向，便于读者选择性学习。

(4) 数字化资源丰富，便于学习。本教材教学资源丰富，读者可通过扫描右侧的二维码获取教学课件、案例数据、R 与 Python 软件代码、习题答案等数字资源，还可通过扫描文中二维码进行在线测试、观看学习视频。已建设完成的省级精品在线开放课程网址，可通过扫描右侧二维码获取。

教学资源

本教材共分为 8 章。第 1 章为概述，主要介绍数据挖掘的发展历程、过程模型、功能、机器学习、应用领域等；第 2 章主要介绍数据与数据平台；第 3 章介绍数据预处理与特征工程；第 4～8 章介绍各类数据挖掘与机器学习方法的基本概念、经典算法及基于 R 和 Python 的实践案例。

本教材主要针对统计学、数据科学与大数据等相关专业的高年级本科生和硕士研究生编写，以帮助学生领悟数据挖掘与机器学习的精髓，掌握从数据中挖掘知识、从模型中获取决策依据的能力，并为其未来在学术研究或行业实践中应用打下坚实基础。本教材也可作为其他数据挖掘与机器学习爱好者的参考用书。

结合笔者近二十年的教学实践，以 48 学时为例(一学期 16 周，每周 3 学时)，本教材

的理论教学内容建议安排 33 学时，第 4～8 章的实践内容建议安排 15 学时。在编写过程中，笔者参考了国内外相关领域许多学者的研究成果，在此深表谢意！

笔者虽已尽心竭力，但限于水平，书中谬误之处在所难免，敬请读者批评指正。

编者

2025 年 7 月于杭州

目　录

第1章

概　　述

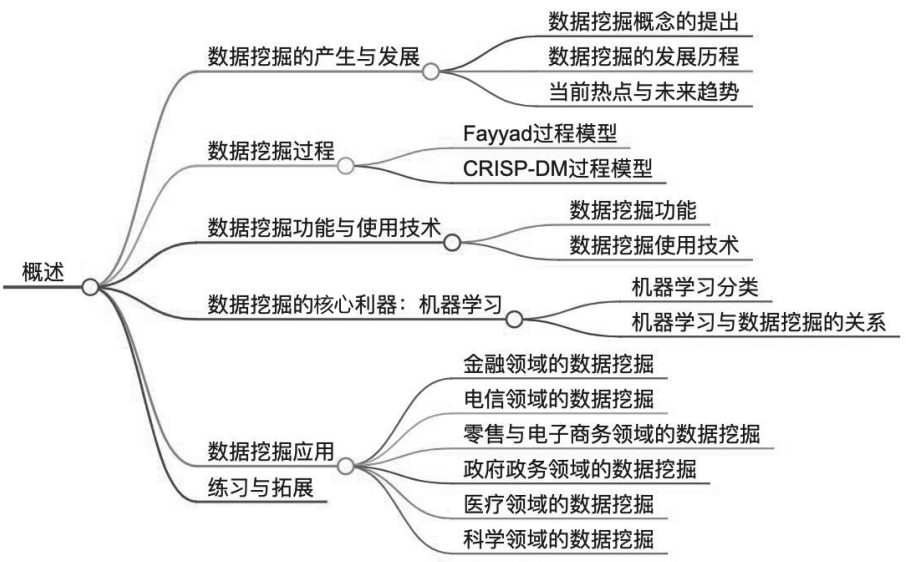
导　读

什么是新质生产力、如何发展新质生产力？我一直在思考，也注意到学术界的一些研究成果。概括地说，新质生产力是创新起主导作用，摆脱传统经济增长方式、生产力发展路径，具有高科技、高效能、高质量特征，符合新发展理念的先进生产力质态。它由技术革命性突破、生产要素创新性配置、产业深度转型升级而催生，以劳动者、劳动资料、劳动对象及其优化组合的跃升为基本内涵，以全要素生产率大幅提升为核心标志，特点是创新，关键在质优，本质是先进生产力。

新质生产力的显著特点是创新，既包括技术和业态模式层面的创新，也包括管理和制度层面的创新。必须继续做好创新这篇大文章，推动新质生产力加快发展。

——摘自习近平 2024 年 1 月 31 日在二十届中央政治局第十一次集体学习时的讲话

知识导图



1.1 数据挖掘的产生与发展

自 20 世纪 60 年代以来，随着信息技术的飞速发展，数据库及数据仓库技术被广泛应用，遍及超级销售市场、银行、天文学研究、医学研究及政府部门等各个领域。以全球最大的零售企业沃尔玛为例，其创始人山姆·沃尔顿非常重视信息的沟通和信息系统的建设，早在 1969 年，便购买第一台计算机用来支持公司日常业务。20 世纪 70 年代，沃尔玛建立了物流的管理信息系统(management information system，MIS)。20 世纪 80 年代初，沃尔玛与休斯公司合作发射物流通信卫星，实现了全球联网；1983 年开始使用 POS 机；1985 年建立了电子数据交换系统(electronic data interchange，EDI)，开始无纸化作业，所有信息都在电脑上运作；1986 年建立了快速反应系统(quick response，QR)，用于订货业务和付款通知业务。20 世纪 90 年代，沃尔玛开始采用全球领先的卫星定位系统(GPS)，控制公司物流。由此，沃尔玛成为全球第一个实现集团内部 24 小时计算机物流网络化监控的企业，实现采购、库存、订货、配送和销售一体化。信息化建设使沃尔玛积累了大量的各类业务数据，但是我们知道，数据作为一种资源，本身并没有什么直接的价值，有价值的是从中所能获得的信息和知识。数据挖掘正是基于这种需要而产生、发展起来的，也由此有了广为流传的“啤酒和尿布”的故事。

据说在 20 世纪 90 年代，沃尔玛对其在美国本土超市的销售数据展开研究，结果发现，和尿布一起购买次数最多的商品竟然是啤酒！啤酒和尿布，似乎风马牛不相及，沃尔玛管理层对这个结果产生了疑问：真是这样吗？为什么？于是，沃尔玛决定对同时购买过啤酒和尿布的客户进行电话回访，询问其为什么会同时购买这两种商品。答案是一些年轻的爸爸在下班途中经常会接到妻子的电话，要求其在回家途中购买孩子的尿布，有 30%～40%的爸爸会顺便买点啤酒犒劳自己。证实了这个规律后，管理层就把啤酒和尿布摆放在一起进行销售，不出意料，销售量双双增加。

1.1.1 数据挖掘概念的提出

1. KDD 国际学术会议

1989 年 8 月在美国底特律召开的第 11 届国际联合人工智能学术会议(IJCAI-89)上，Gregory Piatetsky-Shapiro 组织了“数据库中的知识发现”(KDD：Knowledge Discovery in Database)专题讨论会。该讨论会聚焦于“发现的方法”及“发现的知识”两个方面，这是基于数据挖掘概念的首次国际学术会议。

随后在 1991 年、1993 年和 1994 年都举行了 KDD 专题讨论会，来自各个领域的研究人员和应用开发者集中讨论了数据统计、海量数据分析算法、知识表示和知识运用等问题。随着参与科研和开发人员的不断增加，国际 KDD 组委会于 1995 年把专题讨论会发展成为国际年会。在加拿大的蒙特利尔市召开了第 1 届 KDD 国际学术会议，会议全称为

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining，是世界数据挖掘领域的顶级学术会议。在这次会议上，“数据挖掘”(data mining)概念第一次由 Usama M. Fayyad 提出。Fayyad 同时界定了数据挖掘的内涵，指出数据挖掘是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、有效的、新颖的、潜在有用的并且最终可理解的模式的非平凡过程。以后每年召开一次，参加人数由几十人发展到数千人，研究重点也逐渐从发现方法转向系统应用，并且注重多种发现策略和技术的集成，以及多种学科之间的相互渗透。其中，1997 年第 3 届 KDD 国际学术大会上进行的数据挖掘工具的竞赛评奖活动，就是一个生动的证明。1998 年，在美国纽约举行的第 4 届 KDD 国际学术会议上，与会者不仅进行了学术讨论，而且领略了 30 多家软件公司展示的数据挖掘软件产品。第 31 届 ACM SIGKDD 于 2025 年 8 月 3 日至 7 日在加拿大多伦多举行。

2. 其他国际性数据挖掘年会

除了美国人工智能协会主办的 KDD 年会外，还有许多国际性数据挖掘年会，包括 ICDM、SDM、PAKDD、ECML-PKDD 等。ICDM(IEEE International Conference on Data Mining)是由 IEEE(Institute of Electrical and Electronics Engineers)组织主办的国际数据挖掘会议，会议涉及数据挖掘的所有内容，包括算法、软件、系统及应用，从 2001 年开始，每年召开一次，第 25 届会议于 2025 年 11 月 12 日至 15 日在美国华盛顿举行。SDM(SIAM International Conference on Data Mining)是 SIAM(Society for Industrial and Applied Mathematics)组织召开的数据挖掘讨论会，2001 年 4 月召开第 1 届讨论会，专注于科学数据的数据挖掘，之后每年召开一次，第 25 届会议于 2025 年 5 月 1 日至 3 日在美国弗吉尼亚州的亚历山大市举行。PAKDD(Pacific-Asia Conference on Knowledge Discovery and Data Mining)是亚太地区数据挖掘年会，从 1997 年开始，每年召开一次，第 29 届 PAKDD 于 2025 年 6 月 10 日至 13 日在澳大利亚悉尼举行。PKDD(Principles and Pratice of Knowledge Discovery in Database)是欧洲数据挖掘会议，也是从 1997 年开始，每年召开一次。但是从 2008 年开始，PKDD 已和欧洲机器学习会议(European Conference on Machine Learning，ECML)合并，称为 ECML-PKDD。合并后的 ECML-PKDD 成为欧洲乃至全球范围内机器学习和数据挖掘领域的重要会议，每年吸引大量学术界和工业界的研究人员参与。2025 年 ECML-PKDD 于 9 月 15 日至 19 日在葡萄牙波尔图举行。

1.1.2　数据挖掘的发展历程

数据挖掘技术所表现出的广阔应用前景及其所蕴含的巨大商业价值，吸引了国内外众多研究人员和商业机构从事数据挖掘系统的理论研究和原型开发。

1. 四代数据挖掘系统：基于技术角度的划分

从数据挖掘系统研究的技术角度看，早在 1998 年，Grossman 就提出把数据挖掘系统发展划分为四代的观点，如表 1.1 所示。

表 1.1 四代数据挖掘系统

代	特征	数据挖掘算法	集成	计算模型 分布形式	支持的数据 类型
第一代	独立应用程序	一个或少数几个算法	独立的系统	单台机器	向量数据
第二代	与数据库和数据仓库集成	多个算法；能够挖掘一次不能放进内存的数据	数据管理系统，包括数据库与数据仓库	同质、局部区域的计算机集群	一些系统支持对象、文本和连续的媒体数据
第三代	与预言模型系统集成	多个算法	数据管理系统和预言模型系统	内部/外部网络计算	半结构化数据和 Web 数据
第四代	与移动设备及各种计算设备结合(普适计算)	多个算法	数据管理系统、预言模型系统、移动系统	移动和各种计算设备(普适计算)	普遍存在的各种类型数据

1) 第一代数据挖掘系统

第一代数据挖掘系统支持一个或少数几个数据挖掘算法，这些算法用来支持挖掘向量数据，作为一个独立的系统在单台机器上运行，数据一般一次性调进内存进行处理。这类工具要求用户对具体的算法和数据挖掘技术有相当的了解，还要预先完成大量的数据预处理工作。典型的系统有 Salford Systems 公司早期推出的 CART 系统等。

2) 第二代数据挖掘系统

如果数据量非常大，需要利用数据库与数据仓库技术进行管理，第一代数据挖掘系统显然不能满足需求。第二代数据挖掘系统的主要特点是能够与数据库管理系统(DBMS)集成，支持数据库和数据仓库系统，与它们具有高性能的接口，具有高的可扩展性，支持多个算法，能够挖掘一次不能放进内存的数据，而且有些系统还能够支持挖掘对象、文本和连续的媒体数据。典型的系统如 DBMiner，能通过 DMQL 挖掘语言进行挖掘操作。

3) 第三代数据挖掘系统

第三代数据挖掘系统除了可以与数据管理系统集成外，一个重要的优点是由数据挖掘系统产生的预言模型能够自动地被操作型系统吸收，从而与操作型系统中的预言模型相联合，提供决策支持的功能。另一个特点是支持半结构化数据和 Web 数据，能够挖掘网络环境下的分布式和高度异质的数据，并且能够有效地与操作型系统集成。典型的系统(如早期被 SPSS 公司收购的 Clementine)以 PMML 格式提供与预言模型系统的接口。Clementine 系统现在被命名为 IBM SPSS Modeler，是 IBM 公司的数据挖掘工具之一。

PMML(predictive model markup language)是一种与平台无关的统计和数据挖掘模型表示标准，由数据挖掘协会(Data Mining Group，DMG)开发，已经被 W3C(万维网联盟)接受，成为对数据挖掘模型进行描述和定义的国际标准。PMML 通过定义规范化的数据挖掘建模过程及统一的模型表达，使得模型构造和基于模型的预测功能得以分离并可模块化实现，使得不同平台、不同数据挖掘产品之间能够共享所获得的数据挖掘模型，并为基于模型的

可视化提供了条件。

4) 第四代数据挖掘系统

第四代数据挖掘系统旨在挖掘嵌入式系统、移动系统及各种普适计算设备产生的各种类型数据。普适计算(ubiquitous computing)是软件工程和计算机科学中的概念，指可以使用任何设备，在任何位置，以任何格式进行计算。用户与计算机交互，计算机可以以许多不同的形式存在，包括膝上型计算机、平板电脑和日常生活中的终端，例如汽车、冰箱或一副眼镜。支持普适计算的基础技术包括 Internet、高级中间件、操作系统、移动代码、传感器、微处理器、新的输入输出(I / O)，还包括用户界面、网络、移动协议、位置和定位技术及新材料。物联网的不断发展，云计算、雾计算技术的广泛应用，将会进一步推动第四代数据挖掘系统的研究与发展。

2. 数据挖掘系统发展的三个阶段：基于应用角度的划分

从应用的角度，朱建秋将数据挖掘系统的发展归纳为三个阶段。

1) 独立的数据挖掘系统

独立的数据挖掘系统对应第一代数据挖掘系统，出现在数据挖掘技术发展早期。一般研究人员开发出一种新型的数据挖掘算法，就会形成一个软件，如 1993 年 Quinlan 提出的 C4.5 决策树算法，1994 年 Agrawal 和 Srikant 提出的 Apriori 关联挖掘算法等。

2) 横向的数据挖掘工具

随着数据量的增大，数据库与数据仓库技术广泛应用于数据管理，数据挖掘系统与数据库和数据仓库的结合成为必然的选择；现实领域问题的多样性，导致一种或少数几种数据挖掘算法难以解决所有的问题；用于挖掘的数据通常不符合算法的要求，需要有数据清洗、转换等预处理的配合，才能得出有价值的模型。由于以上三方面的原因，人们认识到数据挖掘软件迫切需要结合数据库和数据仓库、多种类型的数据挖掘算法，以及数据清洗、转换等预处理功能。1995 年前后，软件开发商开始提供称为“工具集”的数据挖掘系统。此类系统的特点是提供多种数据挖掘算法(通常包含分类、聚类和关联等)，同时提供数据的预处理与可视化，是通用算法的集合，并非针对特定的应用，所以称为横向的数据挖掘工具。典型的横向工具有 IBM 公司的 IBM Intelligent Miner、IBM SPSS Modeler 和 SAS 公司的 Enterprise Miner 等。

3) 纵向的数据挖掘解决方案

分析人员使用横向数据挖掘工具不仅需要熟悉分析的业务问题，还要精通数据挖掘算法。如果不了解业务或者算法，就难以获得有效的模型用于决策。从 1999 年开始，国外大量的数据挖掘工具研制者开始提供纵向的数据挖掘解决方案，即针对特定的应用提供完整的数据挖掘方案，如在客户关系管理系统中嵌入基于神经网络的客户流失分析功能；在欺诈防护系统中嵌入基于贝叶斯的欺诈行为预测功能；在零售管理系统中嵌入客户行为分析功能，预测客户购买情况并发送相应的优惠；在机场管理系统中嵌入旅客人数预测功能；在生产制造系统中嵌入质量控制功能等。

1.1.3 当前热点与未来趋势

1. 云计算与大数据

2006 年，谷歌首席执行官埃里克·施密特推出了“Google 101 计划”，正式提出“云”的概念和理论。2008 年 2 月，美国《商业周刊》发表了一篇题为“Google 及其云智慧”的文章，文章开篇就宣称：“这项全新的战略旨在把强大得超乎想象的计算能力分布到众人手中。”随后各大 IT 公司相继推出了自己的“云计划”。中国自 2009 年以来也把“云计算”“云服务”提升到生产方式的高度。国内各大电信企业、地方政府和相关企业先后启动了云计算项目。所有这一切，预示着云计算和大数据时代的到来。

1) 云计算

2006 年，云计算创始人谷歌工程师克里斯托夫·比希利亚向首席执行官埃里克·施密特提出以谷歌设备为核心的“云计算”的想法。谷歌提供在线的网页创建、文档处理、电子表格处理等服务，用户只需要通过网络连接到谷歌的计算“云”，就可以执行相应的操作，而且能实现多人协同工作。自此，业界展开了“什么是云”“什么是云计算”“什么是云服务”的热烈讨论。

Mather 等基于 5 个特性来定义云计算：多重租赁(分享资源)、大规模可扩展性、弹性、随用随付及自行配置资源。Vaquero 等分析了已有关于云计算的定义，认为现有定义都较多地体现某一项技术，缺乏全面性和综合性，其通过界定“云”将云计算定义为：云是一个具有大量易得易用的虚拟资源(如硬件、开发平台或服务)的资源池，这些资源可以根据不同的需求规模进行动态的重新分配，以提高资源的利用率，并实行按使用量付费的支付模式。Wang 等从云计算系统功能的角度给出了云计算系统的定义，指出云计算系统不仅能向用户提供硬件即服务(hardware as a service，HaaS)、软件即服务(software as a service，SaaS)、数据资源即服务(data as a service，DaaS)，还能够向用户提供能够配置的平台即服务(platform as a service，PaaS)，因此用户可以按需向计算平台提交自己的硬件配置、软件安装、数据访问需求。Fingar 认为“云”包含三个层面：①云计算，即一种设计模式，可实现自助服务自动化、可扩展、灵活、费用机动、数据分析方法丰富多样；②云平台，即各种工具、编程与信息模型、辅助软件运行的组件及相关技术；③云服务，即一种用于信息服务的分发模型。Armbrust 等认为云计算既指在互联网上以服务形式提供的应用，也指在数据中心里提供这些服务的硬件和软件，而这些硬件和软件被称为“云”。姚宏宇和田溯宁认为云计算应该包括服务和平台两方面内容，云计算既是商业模式，也是技术。

基于以上不同学者的分析，本书认为云计算不仅是技术，更是一种全新的商业服务模式。云计算服务以云资源为实现基础，以云计算技术为实现保障，以低成本、按需付费的形式，向用户提供软(硬)件基础设施、计算平台和软件服务，使用户在无基础投入的前提下直接实现数据的存储、管理和分析，也可利用提供的云服务平台创建和开发应用程序，或者直接使用云服务平台提供的各类服务软件。

2) 大数据

对于大数据，虽然众说纷纭，但有一个相对一致的说法是：大数据是超出了典型(传统、常用)硬件环境和软件工具收集、存储、管理和分析能力的数据集。由此可知，“大数据”是一个动态发展的、相对的概念。随着软(硬)件技术的发展，大数据的内涵会发生相应的变化。结合目前常用的软(硬)件技术，当下的“大数据”可以具体理解为日常关系型数据库无法收集、存储和管理的数据集。关系型数据库适合管理结构化数据，所以，当下的“大数据”除了数据量庞大(一般指 PB 量级及以上)，数据形式还复杂、多样，不仅有大量的结构化数据，还有大量半结构化及非结构化的数据。社交网站、智能化移动设备及传感器的大规模使用，促使数据产生的速度越来越快，半结构化和非结构化的数据已占据主导地位。虽然因为数据量大，数据的价值密度较低，但从绝对数量看，大数据中蕴含着大量有价值的信息。

正是因为大数据中蕴含着大量有价值的信息，大数据被人们认为是下一个社会发展阶段的石油和金矿。各个国家把大数据当作一种全新的社会资源，并把大数据产业的发展提升到国家战略发展的高度。石油的勘探、开采、运输、提炼与石油产品的生产与销售等多个环节构成了石油产业，类比于石油资源，大数据的生产、采集、传输、存储、分析及应用则构成了大数据产业。在大数据产业链中，大数据分析环节非常重要。它既是前几个环节的成果体现，又是大数据应用及创新的基础。大数据分析的需要促进了大数据挖掘的发展，与传统的数据挖掘相比，大数据挖掘将更多依赖于云计算技术，虚拟化、可扩展的分布式数据存储模式使数据存储不仅在量上没有了限制，而且数据形式也更复杂，不仅包含了大量半结构化和非结构化的数据，还包括大量流数据。大数据挖掘将面临更海量的数据，更复杂的数据预处理过程，更多变的挖掘环境。

随着人工智能、云计算和大数据技术的进步，数据挖掘应用领域不断拓展。以下是当前数据挖掘的主要热点及未来发展趋势。

2. 当前热点

1) 多模态数据挖掘

多模态数据挖掘是指从多种类型的数据(如文本、图像、音频、视频等)中提取有价值的信息和知识，以提高数据挖掘的有效性和应用范围。随着多媒体数据的快速增长，多模态数据挖掘成为数据挖掘领域的一个重要研究方向。多模态数据挖掘涉及多个核心技术，主要包括多模态数据表示、多模态数据融合、跨模态对齐(时间对齐、语义对齐等)、多模态数据挖掘模型(如 CNN+LSTM：用于视频+音频分析)等。

多模态数据挖掘在多个领域有广泛应用，如结合 CT、MRI 影像和病人文本病历进行数据挖掘，以提高疾病诊断准确率；融合摄像头、雷达和 GPS 数据，以提高车辆环境感知能力，优化自动驾驶决策；结合音乐、视频与用户行为数据，以优化音乐或短视频推荐算法；结合监控视频和环境声音检测异常行为；等等。

尽管面临数据融合、语义对齐等挑战，但随着深度学习、联邦学习等技术的发展，多模态数据挖掘将成为未来 AI 的重要方向。

2) 实时流数据挖掘

实时流数据挖掘指的是从持续产生的数据流(如金融交易、传感器数据、社交媒体、网络日志等)中动态挖掘有价值的信息，并在低延迟的情况下进行实时分析，以便快速响应决策需求。随着物联网、社交媒体和金融交易等领域的快速发展，实时流数据挖掘成为数据挖掘领域的一个重要研究方向。实时流数据挖掘涉及多个核心技术，包括流式计算架构、流数据处理算法、存储优化等。

实时流数据挖掘在金融交易、物联网、社交媒体、网络安全、交通管理等领域的应用前景非常广阔，如实时监控金融交易数据，检测异常交易和欺诈行为；实时分析智能家居设备数据，提供个性化服务；实时分析社交媒体数据，识别用户情感和舆论趋势；实时分析系统日志，识别安全威胁和异常行为；实时监控交通流量数据，优化交通管理和调度；等等。

3. 未来趋势

1) 数据挖掘与各专业领域持续深入结合

早在 2011 年，全球知名咨询管理公司麦肯锡在其一份研究报告《大数据：下一个创新、竞争和生产力的前沿》中提出：“数据，已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。人们对于海量数据的挖掘和运用，预示着新一波生产率增长和消费者盈余浪潮的到来。”各行业的生产系统每时每刻都在产生海量数据，如政务管理数据、电子商务数据、物联网传感器数据、医疗数据等。与各行业生产系统的深度结合，对这些数据展开广泛深入的挖掘，不仅可以推动这些行业向前发展，也是数据挖掘保持长久生命力的源泉。

2) 数据挖掘与 AI 在应用层面的不断融合

大数据技术的加速发展，使得从海量数据中获取智能成为可能。数据挖掘技术，尤其是作为其技术支撑之一的机器学习方法，将在未来各类应用系统(例如智慧城市、智慧医疗、智慧交通、智慧家居等)中，与 AI 不断融合，共同发展。

3) 数据挖掘与云计算、边缘计算的紧密结合

目前，很多人认为，云计算是解决大数据生产、采集、传输、存储、分析及应用的最佳平台之一。人们在提到大数据的时候，总会想到云计算。云计算强调的是技术，大数据强调的是效用和价值。数据规模持续呈指数级增长，本地存储和计算能力有限，云计算提供近乎无限的弹性扩展能力，支持海量数据存储与处理。所以，未来数据挖掘与云计算的结合将更加紧密。

边缘计算是一种分布式计算范式，将计算资源和数据处理能力推向网络的边缘，靠近数据源。与云计算模式相比，边缘计算能够减少数据传输延迟，提高实时性和响应速度，适用于物联网、智能交通、工业制造等领域。随着低延迟需求的增加、隐私保护要求的提升，以及 AI 技术的发展，未来数据挖掘与边缘计算的结合也将更加紧密。

4) 数据挖掘与区块链技术的逐步结合

区块链技术被认为是互联网发明以来最具有颠覆性的技术创新之一，它依靠分布式算法，不依赖任何第三方中心，通过自身分布式节点进行网络数据的存储、验证、传递和交

流。区块链的不可篡改性确保数据的真实性和完整性，可提高数据挖掘结果的可信度。利用区块链的智能合约功能，可自动完成数据挖掘任务的执行和管理工作。区块链技术通过加密和匿名化手段，支持在数据挖掘过程中保护用户隐私。未来，随着跨链技术和绿色计算等技术的发展，数据挖掘与区块链的结合将更加紧密，推动更多创新应用的出现。

1.2　数据挖掘过程

从工程学的角度来看，数据挖掘是一个多环节、多处理阶段的闭环过程。如同软件工程中的软件过程模型在软件开发中的作用，数据挖掘过程模型为数据挖掘提供了宏观指导和工程方法。早期人们进行数据挖掘研究是为了将发现的研究成果应用于实际数据处理中，为科学决策提供支持。因此，大多数研究人员只着眼于数据挖掘的算法和应用层面，而忽视了其他方面。事实上，数据挖掘首先是一个处理过程，如果我们仅仅着重于挖掘，可能就看不到实际数据处理过程中数据提取、组织和显示的难度。合理的数据挖掘过程模型能将各个处理阶段有机地结合在一起，指导人们更好地开发、使用数据挖掘系统和实施数据挖掘项目。从数据挖掘进入工程应用领域起，就有人对数据挖掘的过程进行归纳和总结，以便人们开发及使用数据挖掘应用系统。目前，被业界广泛认可并已应用于商用软件的数据挖掘过程模型主要有两种：一种是 Fayyad 等人总结的过程模型，另一种是遵循 CRISP-DM 标准的过程模型。

1.2.1　Fayyad 过程模型

Fayyad 等将知识发现过程定义为：从数据中鉴别出有效模式的非平凡过程，该模式是新颖的、可能有用的和最终可理解的。图 1.1 是 Fayyad 过程模型。早期开发的大部分数据挖掘系统都是遵循 Fayyad 过程模型，例如 IBM Intelligent Miner 和 SAS Enterprise Miner 等。

如图 1.1 所示，Fayyad 过程模型包括数据选择(data selection)、数据预处理(data preprocessing)、数据转换(data transformation)、数据挖掘(data mining)、模式解释与评价(pattern interpretation and evalution)。

1. 数据选择

数据选择是指根据分析任务的要求从原始数据中提取和挖掘与目标相关的数据，并将不同数据源中的数据集成在一起，形成本次数据挖掘任务的数据集。在此过程中，会利用一些数据库操作对数据进行处理。

2. 数据预处理

数据预处理是指对数据选择阶段产生的数据进行再加工，检查数据的完整性及数据的一致性，对其中的噪声数据进行处理，对缺失的数据进行填补等。

3. 数据转换

数据转换是指对经过预处理的数据，根据挖掘事务的任务对数据进行再处理，主要是将其转换成数据挖掘算法所需要的形式，如将连续型数据转换成离散型数据等。

4. 数据挖掘

数据挖掘是指运用合适的数据挖掘算法，从数据中提取出用户所需要的知识，这些知识可以用一种特定的方式表示或使用一些常用的表示方式，如产生规则等。

5. 模式解释与评价

模式解释与评价是指根据分析目的，对发现的模式进行解释，并评价模式的有效性。在此过程中，为了取得更有效的模式，可能会返回到前面的某些处理步骤，从而提取出更有用的知识。

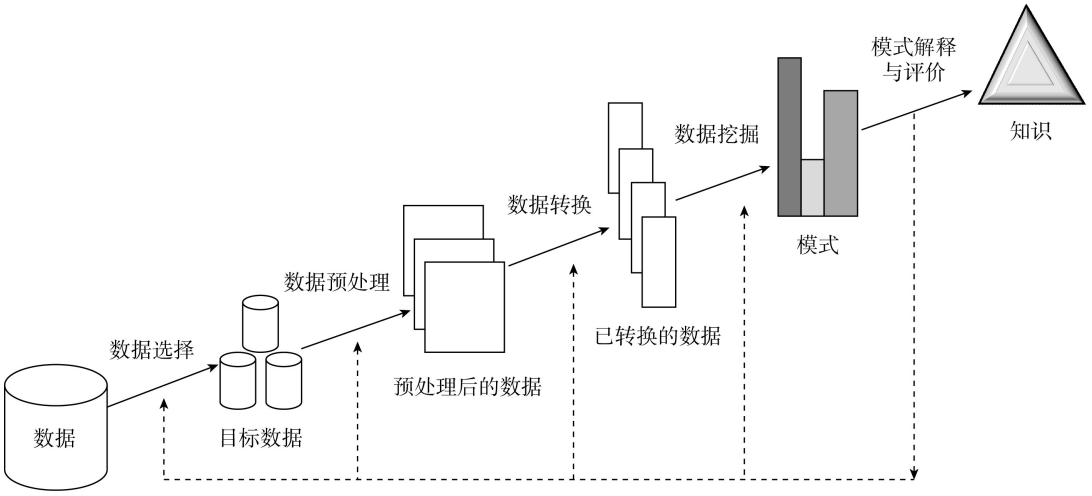


图 1.1　Fayyad 过程模型

从上述 Fayyad 过程模型看，这个过程已经包括了数据挖掘过程中各个必要的处理阶段，并且形成了一个可以根据各个处理阶段的结果来决定是否返回以前的阶段进行再处理的闭环过程。但是，Fayyad 过程模型从数据入手，到知识结束，过多地偏重从技术的角度来理解数据挖掘过程。在实际使用过程中会存在两个问题：①数据选择对于整个分析至关重要，但是该如何选择，选择哪些数据呢？这是由具体的商业问题决定的，需要领域专家、数据管理员与数据挖掘专家一起讨论确定。如何明确商业问题，并把商业问题和数据相关联，这在 Fayyad 过程模型中没有反映。②数据挖掘一般在分析型环境中获得知识，获得的知识只有返回到操作型环境中使用，才能产生真正的价值。在 Fayyad 过程模型中，模式评价阶段结束后，对于挖掘到的知识应该如何使用，也没有体现。

1.2.2　CRISP-DM 过程模型

CRISP-DM(cross-industry standard process for data mining，跨行业数据挖掘标准过程)

由 SPSS、NCR 及当时的戴姆勒-克莱斯勒等公司在 1996 年提出，后来得到欧洲共同体研究基金的资助。2000 年 8 月，CRISP-DM 1.0 版正式推出。CRISP-DM 强调，数据挖掘不单是数据的组织或者呈现，也不仅是数据分析和统计建模，而是一个从理解业务需求、寻求解决方案到接受实践检验的完整过程。如图 1.2 所示，CRISP-DM 过程模型包括商业理解(business understanding)、数据理解(data understanding)、数据准备(data preparation)、建模(modeling)、评价(evaluation)和部署(deployment)6 个阶段。图 1.2 的外圈形象地表达了数据挖掘过程的循环特性。通常，一个数据挖掘项目并不是一次部署完就结束，在挖掘的过程中或部署过程中获得的经验可能会触发新的商业问题，后续的挖掘过程将从前一次的经验中受益，并做出相应的调整。内部的箭头表示阶段之间最重要和最频繁发生的关联关系。阶段间的顺序不是严格不变的，可以根据具体任务的需要进行来回选择。

CRISP-DM 不仅被许多数据挖掘软件商用来指导开发数据挖掘软件(如 IBM 公司的 IBM SPSS Modeler 就遵循了 CRISP-DM)，也被广泛用来指导数据挖掘项目的实施。

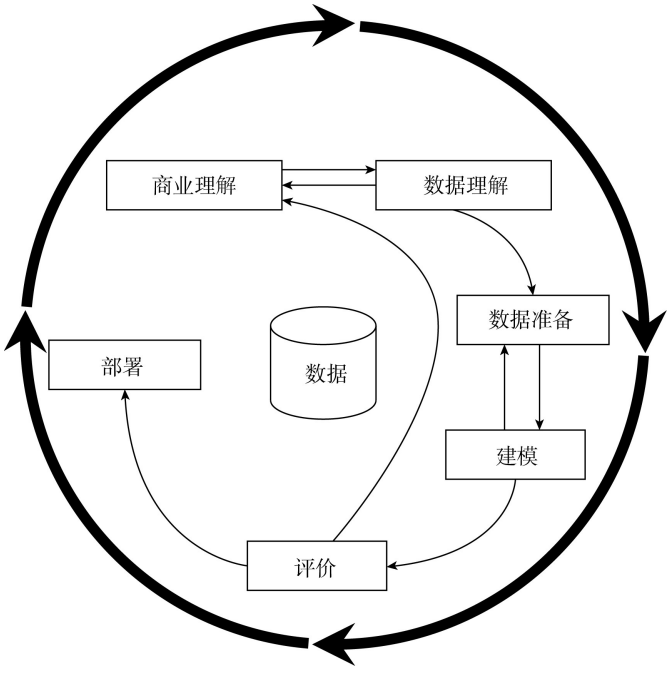


图 1.2　CRISP-DM 过程模型

1. 商业理解

商业理解是对企业运作、业务流程和行业背景进行了解，专注于从商业的角度理解项目目标和需求，然后将这种目标和需求转换成一个数据挖掘的问题及相应的项目计划，其一般任务和输出内容如图 1.3 所示。

1) 确定商业目标

数据分析师最重要的能力是对业务的理解和把握。如果没有正确地理解业务，再好的理论，再强的工具，都只会徒劳无益。所以，一个数据挖掘项目的实施，其首要任务就是从业务的角度真正理解所要解决的问题和所要实现的目标。

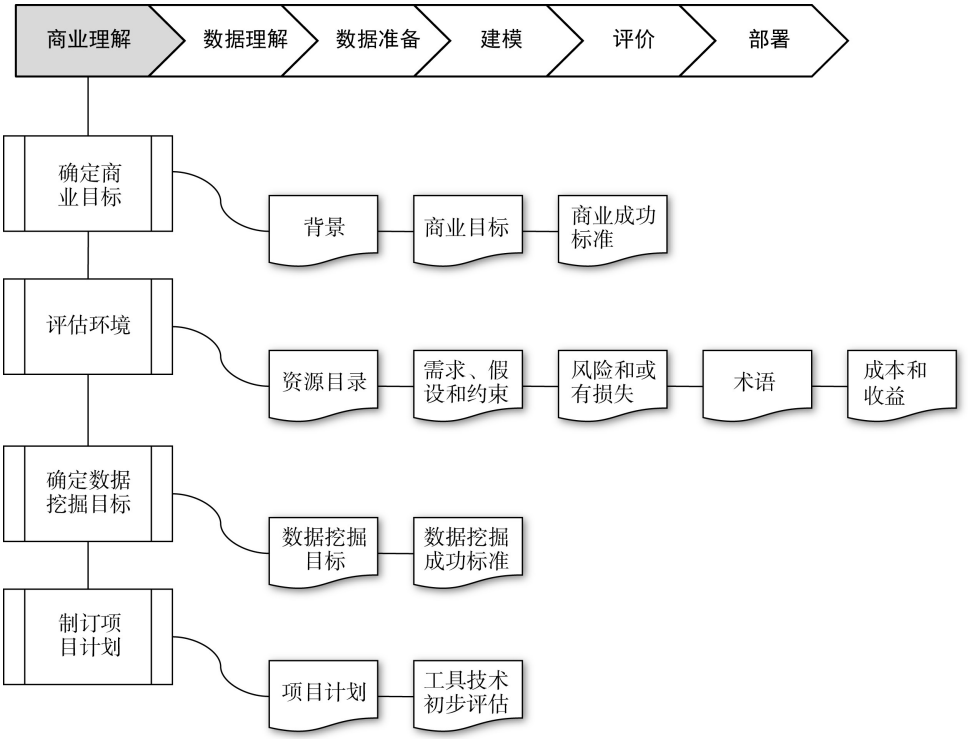


图 1.3　商业理解的一般任务和输出内容

完成确定商业目标这一任务，其相应的输出文档内容一般包括背景、商业目标和商业成功标准三个方面。

(1) 背景包括项目的商业环境，问题涉及的范围，项目的前提(如现有解决方案的优缺点、项目的动机、是否已经使用数据挖掘等)，项目需要的人力和物资，项目将影响到的部门和使用项目结果的目标群体等。

(2) 商业目标是从商业的角度来描述打算用数据挖掘来解决的问题。尽可能准确地分析所有相关的商业问题，分清主要的商业目标及其他次要目标，制定尽可能实现的目标，并使用商业术语，详细说明期望收益。

(3) 商业成功标准是从商业角度衡量项目结果成功的度量标准，包括客观度量标准(如投诉率下降 15%、下单转换率增加 20%等)和主观度量标准。主观度量标准要明确主观的主体，即是谁给出的主观判断。

2) 评估环境

评估环境任务主要围绕已确定的商定目标和初步计划细化各种影响因素，其相应的输出文档内容一般包括资源目录，需求、假设和约束，风险和或有损失，术语，以及成本和收益 5 个方面。

资源目录文档需要列出项目可用的各类资源，包括参与人员(项目发起人、相关商业领域专家、数据库管理员、市场分析师、数据挖掘专家及其他技术支持人员)，数据(企业内部固定抽取的数据、访问内部数据库或数据仓库的数据、外部调查或购买的数据等)，计算资源(硬件平台)和软件(数据挖掘工具及其他相关软件)。

需求、假设和约束文档要求列出项目执行的全部需求、围绕项目整个过程的各方面假设及约束。全部需求可包括：项目完成的时间进度表及相应进度的需求，项目和模型的可理解性、准确性、可部署性、可维护性和可重复性等方面的需求，安全、隐私及法律限制等方面的需求。假设包括对外部因素(如商业环境、经济问题、技术因素等)的假设，数据质量(如可用性、准确度等)的假设，模型理解、解释与评估时可能的假设等。约束包括一般性约束(如法律问题、经费、时间及其他所需资源)，数据源访问权利，数据访问时的技术性问题等。

风险和或有损失(contingencies)文档要求列出可能导致项目延期或失败的风险、可能的损失和为避免这些风险可采取的相应措施。确定每个风险可能发生的条件，如法律风险、商业风险、组织风险、经济风险、技术风险及与数据或数据源有关的风险(数据质量相关问题)等，并计算相应的可能损失，制订损失计划。

术语文档要求编辑一个与项目有关的术语表。术语表至少包括与商业问题有关的术语和与数据挖掘有关的术语两部分内容，以帮助不同专业背景的项目参与人员更好地理解项目。

成本和收益文档要求分析项目执行的成本和项目部署后可能产生的收益(如投资回报率、客户满意度等)。除了数据收集、项目开发和运行等成本，还必须考虑数据重复抽取和准备、工作流程的改变等隐含成本。

3) 确定数据挖掘目标

确定数据挖掘目标这一任务就是要根据已确定的商业目标，从数据挖掘的角度，用数据挖掘技术术语来描述项目目标和项目成功的标准。相应的输出文档内容一般包括数据挖掘目标和数据挖掘成功标准两个方面。

数据挖掘目标要求把商业问题转换成数据挖掘问题，即确定业务问题需要用什么类型的挖掘模型加以解决。若商业目标是要确定哪些客户会流失，则数据挖掘目标是构建一个客户流失预测模型，可以是客户是否流失的分类预测，也可以是客户流失概率预测。

数据挖掘成功标准指模型评估的标准。例如，对于客户是否流失的分类预测模型，可以使用准确率、精准率和召回率等评价指标来评估模型。如果是主观评价标准，和商业成功主观标准一样，需要明确这个标准是由哪个人或哪些人做出的主观判断。

4) 制订项目计划

为达到数据挖掘目标进而实现商业目标，需要制订详细的项目计划。该计划要求详细列出项目需要完成的一系列步骤，包括对工具和技术的选择。相应的输出文档内容一般包括项目计划及工具和技术的初步评估。

项目计划需要列出每个阶段的详细计划，包括持续的时间、需要的资源、输入、输出、可能的风险及关联性。在计划中要交代清楚可能的重复步骤及所需的时间。在估计项目时间进度时可以参考他人的经验，如数据理解和数据准备通常需要占用 60%～80%的时间。应分析时间进度和可能的风险之间的关联性，尽可能避免风险。

工具和技术的选择可能影响整个项目，所以要尽早列出工具和技术的选择标准，评估技术的合适程度，选择最合适的工具和技术。

2. 数据理解

数据理解是对企业现有应用系统进行了解，对数据挖掘所需数据进行全面调查以获取完成挖掘目标所需的初步数据，然后从总体上对获得的数据的属性进行描述，包括数据格式、数据量、一致性、数据出处、收集时间频度等多个方面，并检查数据是否能够满足相关的要求，探索数据和检验数据质量等。其一般任务和输出文档内容如图 1.4 所示。

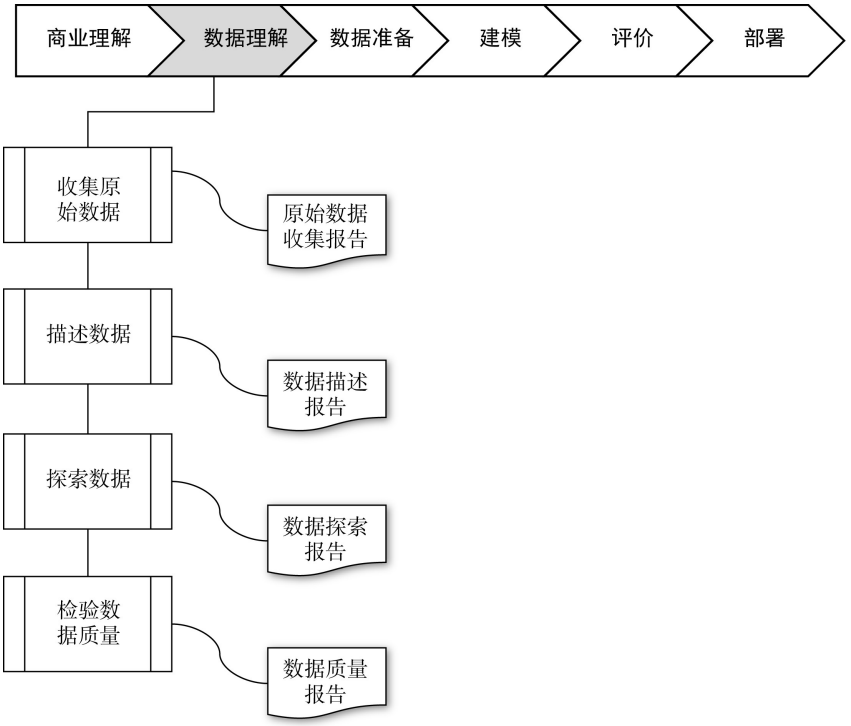


图 1.4　数据理解的一般任务和输出文档内容

1) 收集原始数据

收集原始数据任务是根据资源目录列出的数据资源选择感兴趣的表或文件，并选择表或文件中感兴趣的数据。完成这一任务要求生成相应的输出文档——原始数据收集报告。该报告应包括以下内容：数据来源(内部数据库或数据仓库、外部提供者)，负责维护、收集或购买此数据的人，调查或购买数据需要的费用，数据存储方式，安全和隐私需求、使用限制等。

2) 描述数据

描述数据任务要求描述所获得的数据，包括数据数量(表、各个表的字段数和记录总数)，数据类型，编码方案，计量单位，取值范围或个数，属性和属性值的含义，主键和外键的关系，缺失数据占比等。该任务对应的输出文档是数据描述报告。

3) 探索数据

探索数据任务是根据数据挖掘目标，结合数据描述报告，采用表格、图形和其他可视化技术细致探索数据，包括关键属性的分布、属性间的关系及一些简单的统计分析。这些分析丰富或细化了数据描述，可以作为后续数据准备工作的输入，或者可能直接达到某个

数据挖掘目标。这一任务将生成相应的输出文档——数据探索报告。

4) 检验数据质量

检验数据质量任务需要对收集的数据从是否完整、是否缺失、是否一致、有无异常等方面进行检查，并生成该任务相应的输出文档——数据质量报告。该报告要求列出数据质量检验的结果，对于存在的质量问题，列出可能的解决方法。质量问题的解决方法很大程度上依赖于数据和商业知识。

3. 数据准备

数据准备是数据挖掘过程中最重要的一个环节之一，通常需要耗费大量的时间，一般占用整个数据挖掘项目 50%～70%的时间和工作量。数据准备需要从所收集的大量原始数据中取出一个与业务目标相关的样本数据集，对该数据集进行描述，在此基础上，将该数据集转化为适合数据挖掘工具处理的最终目标数据，包括选择数据、清洗数据、构造数据、集成数据和格式化数据。其一般任务和输出文档内容如图 1.5 所示。

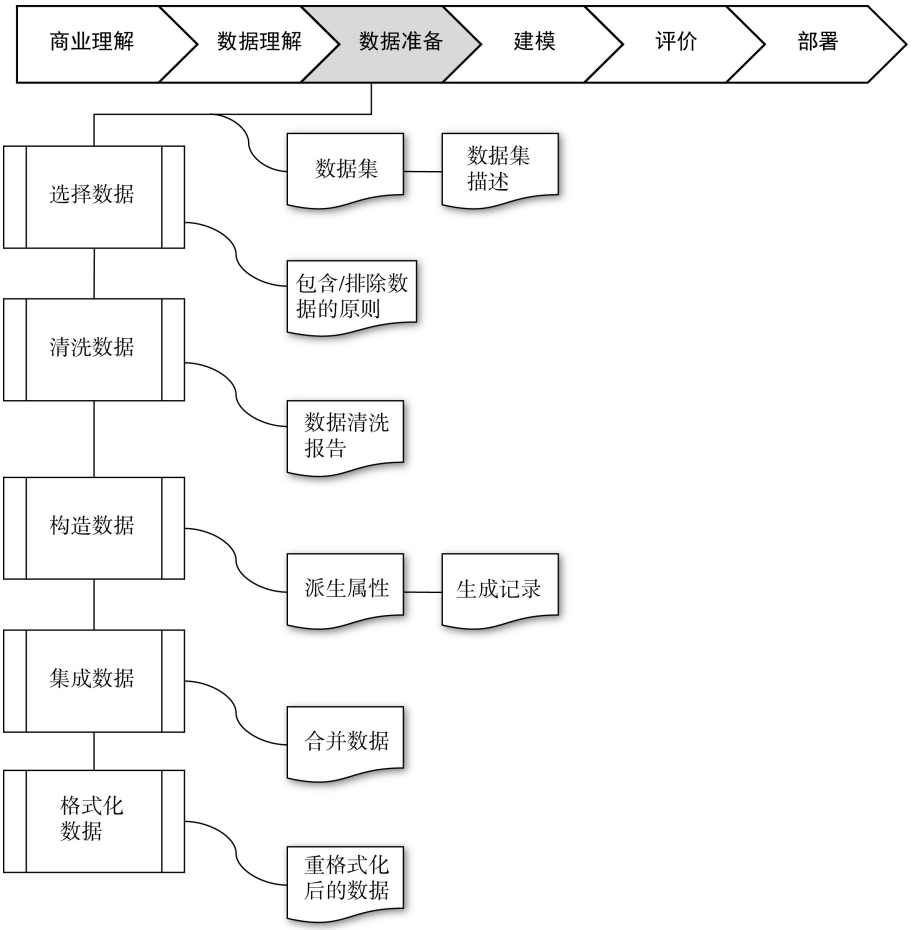


图 1.5　数据准备的一般任务和输出文档内容

1) 选择数据

选择数据需要确定用于分析的数据，包括对样本的选择和对属性或特征的选择。选择的标准直接影响用于分析的数据质量，所以选择标准的确定至关重要，可以从与数据挖掘目标的相关性角度考虑，进行显著性检验或相关性分析，将其作为属性或特征的选择标准，也可以从数据质量、容量与类型等方面限制，将其作为选择的标准。该任务相应的输出文档为包含/排除数据的原则，需要列出被包含进来的和被排除出去的数据，并给出理由。

2) 清洗数据

清洗数据主要是基于已选择的数据，选择合适的方法处理噪声、填补缺失值等，保证数据的正确性和一致性，提升数据质量。其相应的输出文档为数据清洗报告。该报告不仅要描述清洗的策略和行为，还要指出清洗后的数据用于挖掘时仍然可能存在的质量问题及对挖掘结果的潜在影响。

3) 构造数据

构造数据主要指派生属性(列或特征)、生成全新的记录(行)及对现有属性值进行转换等。派生属性是在一个或多个现有属性基础上构造符合挖掘目标需要的属性，例如为了预测客户是否会流失，通过对客户消费行为的分析，界定流失的内涵，构造新的属性“是否流失”，作为目标变量用于预测。该任务相应的输出文档即为构造的结果——派生属性和生成记录。

4) 集成数据

集成数据是指把来自不同数据源的数据整合在一起，可以合并多个表，也可以通过数据合并构造新的记录和属性。例如，一家电子商务公司有两张客户信息表：一张为客户基本信息表，包括客户 ID 号、姓名、年龄、性别等客户基本信息；另一张为客户购买信息表，包括客户 ID 号、客户近一个月的购买明细记录，每一条记录对应每笔购买信息。对这两张表进行集成，可以先根据客户购买信息表生成一个新表，其中每条记录对应每个客户，属性则为客户 ID 号、购买次数、平均购买额、购买促销商品的比例等，再利用客户 ID 号，集成新表和客户信息表。该任务相应的输出文档即为集成的结果——合并数据。

5) 格式化数据

格式化数据作为建模前的最后一个步骤，主要是针对某些建模对数据的特殊格式要求进行调整。例如有些建模算法要求记录按某个属性值排序，有些建模算法又要求记录是随机排列的。对于文本数据，某些建模算法要求去掉文本字段内的标点符号，或者规定每个字段的值所允许的最大字符数。该任务相应的输出文档即为格式化后的结果——重格式化后的数据。

4. 建模

建模是根据对业务目标的理解，在数据准备的基础上，选择和应用多种不同的建模技术，调整它们的参数使其达到最优值，包括选择建模技术、生成测试设计、构建模型和评估模型。其一般任务和输出文档内容如图 1.6 所示。

1) 选择建模技术

选择建模技术是结合数据挖掘目标确定实际所要使用的建模技术，可以是一种技术，

也可以是多种技术，或者是基于多种技术的集成。确定了相应的技术后，需要了解所选技术对数据的假定要求，并产生相应的输出文档——建模技术和建模假设。

2) 生成测试设计

生成测试设计是指在实际构建模型前，建立一个用来测试模型质量和有效性的机制，包括数据集如何划分、划分成几部分(如训练集和测试集)、如何验证模型质量。其相应的输出文档是测试设计。

3) 构建模型

构建模型是指在准备好的数据集上使用建模工具，创建一个或多个模型。相应的输出文档为参数设置、模型和模型描述。参数设置列出模型需要调整的参数、相应的设置值及选择设置值的基本原则。模型是指产生的实际模型，如决策树模型、神经网络模型。模型描述是指描述模型的特征，生成解释模型的报告。

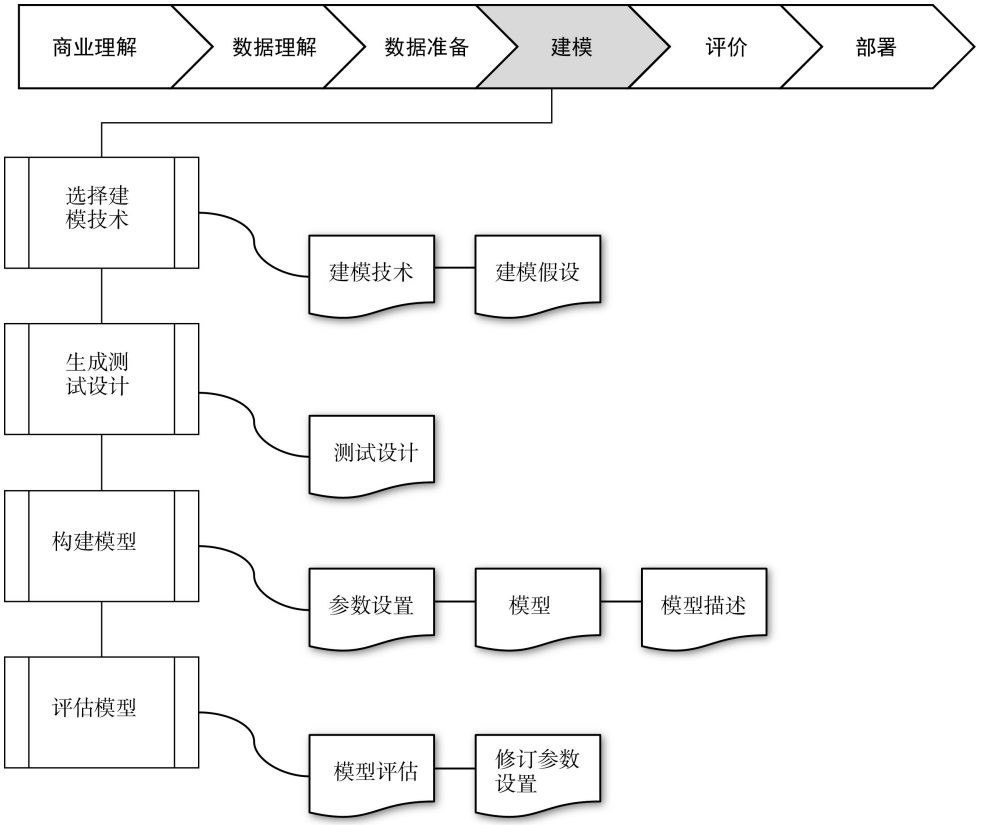


图 1.6　建模的一般任务和输出文档内容

4) 评估模型

评估模型是指数据挖掘工程师根据领域知识、数据挖掘目标成功标准和已生成的测试设计来解释模型。这一任务仅考虑模型，对后续的评价阶段会产生影响。评价阶段需要数据挖掘工程师和领域专家、业务分析人员一起考虑项目实施过程中生成的所有结果。相应的输出文档是模型评估和修订参数设置。模型评估列出全部建成的模型及其评估结果，如

按准确率比较建成模型的优劣。根据模型评估结果，重新修订参数设置，并调整其值建立新的模型，直到数据挖掘工程师确信已找到最优模型为止。修订参数设置指记录所有这些修订和评估。

5. 评价

评价是由分析人员和领域专家一起从业务目标的角度全面地评价得到的模型，以确定它是否完全达到了业务目标，最终做出是否应用数据挖掘结果的决策。其一般任务和输出文档内容如图 1.7 所示。

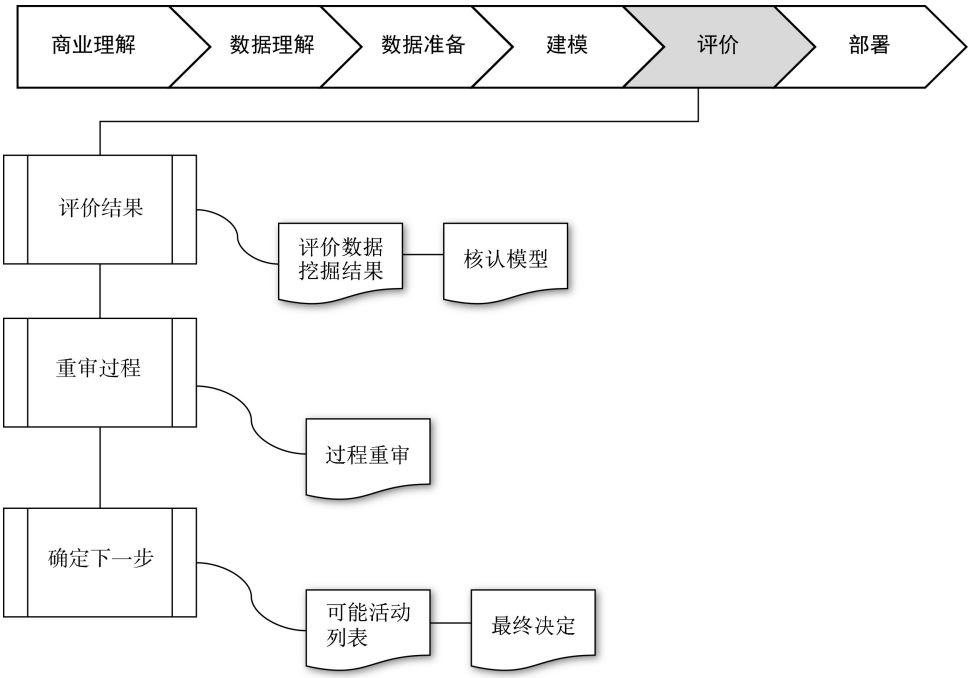


图 1.7　评价的一般任务和输出文档内容

1) 评价结果

评价结果是评价模型是否符合商业目标，若存在不足之处，说明其商业理由，相应的输出文档为评价数据挖掘结果和核认模型。评价数据挖掘结果是指使用商业成功标准术语概述模型评价的结果，包括是否已满足既定商业目标的最终声明。核认模型是核准认可满足既定商业成功标准的模型。

2) 重审过程

重审过程是指对数据挖掘项目实施的整个过程进行重新审核，用来确定是否忽略了某些重要的因素或任务，或者是否存在某些质量问题。其相应的输出文档为过程重审，即概述重审过程，并特别注明被忽略的因素或应该重复的环节。

3) 确定下一步

确定下一步是指根据评价结果和重审过程，来分析项目该如何推进，需要确定是进入部署阶段还是继续重复前面步骤或者创建新的数据挖掘项目，同时，要分析剩余的资源和

预算。其相应的输出文档是可能活动列表和最终决定。可能活动列表列出潜在的进一步活动，并给出支持和反对每个结果的理由。最终决定描述如何合理推进。

6. 部署

部署是数据挖掘的最终目的，是将数据挖掘结果部署到商业环境中，成为日常商业运作的一部分，并生成一份基于项目整个过程的最终报告。其一般任务和输出文档内容如图 1.8 所示。

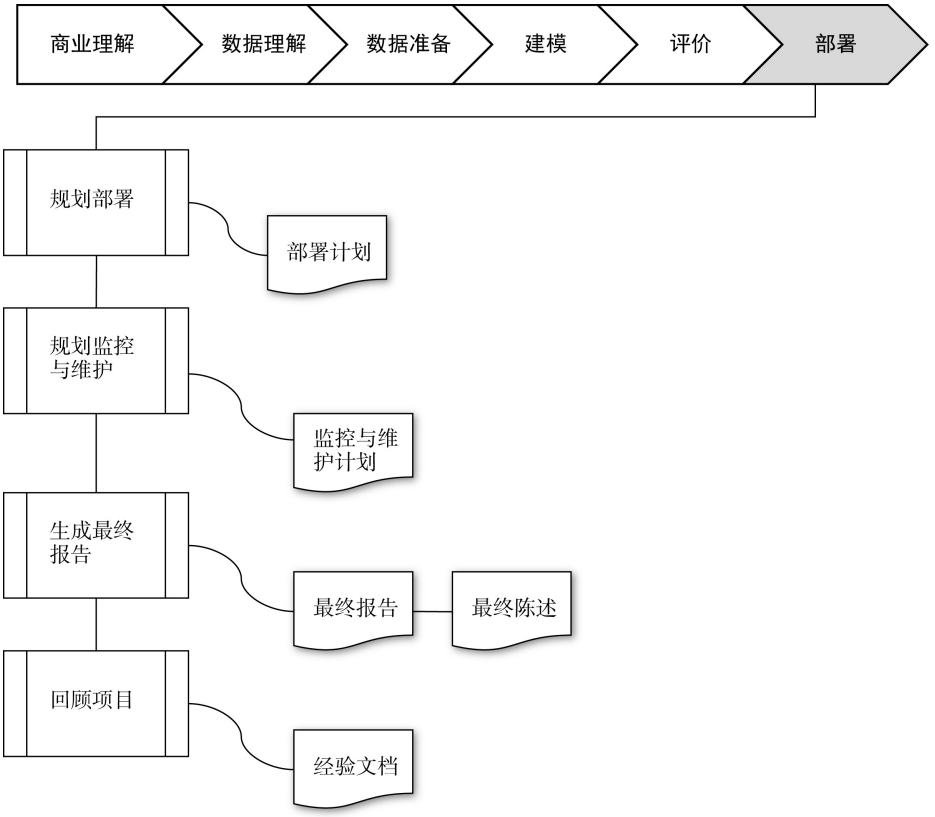


图 1.8　部署的一般任务和输出文档内容

1) 规划部署

规划部署是指为了把数据挖掘结果部署到商业环境中，利用评估的结果给出部署的策略。其相应的输出文档是部署计划，即概述部署策略，包括必要的步骤和如何执行这些步骤。

2) 规划监控与维护

数据挖掘结果成为日常商业运作的一部分时，监控和维护就成为重要问题。规划详细有效的监控和维护策略有助于避免长期错误应用数据挖掘结果。其相应的输出文档是监控与维护计划，即概述监控和维护策略，包括必要的步骤和如何执行这些步骤。

3) 生成最终报告

项目的结束需要项目成员撰写一份最终报告，这份报告可能仅对项目和其经历进行概

述，也可能对数据挖掘结果进行全面展示。其相应的输出文档是最终报告和最终陈述。最终报告可以描述全部过程并标明全部取得的结果，说明与原始计划的偏差，并给出将来工作的建议。其具体内容和形式很大程度依赖于报告的接受者。最终陈述一般只包括最终报告的一部分内容，可以不同于报告的形式呈现。

4) 回顾项目

回顾项目指总结经验，评论成功与失败之处，并指出如何改进。其相应的输出文档为经验文档，即描述项目期间获得的重要经验。

1.3 数据挖掘功能与使用技术

数据挖掘功能用于指定数据挖掘任务发现的模式。一般而言，这些任务可以分为两类：描述性数据挖掘任务和预测性数据挖掘任务。描述性数据挖掘任务是刻画目标数据中数据的一般性质。预测性数据挖掘任务是在当前数据上进行归纳，以便作出预测。随着信息技术的持续发展，数据挖掘吸纳了统计学、机器学习、模式识别、数据库与数据仓库、信息检索、可视化、分布式并行计算等领域的大量技术。

1.3.1 数据挖掘功能

常见的数据挖掘功能可以概括为 6 个方面：数据描述、聚类、偏差检测(孤立点检测)、关联分析、预测和分类，如图 1.9 所示。其中，数据描述、聚类、偏差检测和关联分析可以认为是描述性任务，分类和预测可以认为是预测性任务。

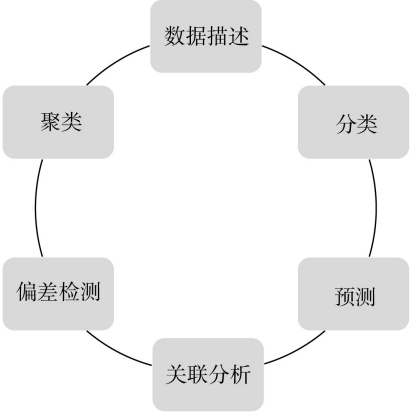


图 1.9 数据挖掘的主要功能

1. 数据描述

数据描述可以分为特征性描述和区别性描述。特征性描述用来反映目标数据的一般特征；区别性描述用来比较目标数据与一个或多个类比数据的不同特征。数据描述通常以图形、二维或多维表的形式呈现描述结果，也可以规则的形式呈现。

2. 聚类

聚类指按照尽量使同一个类(簇)中的数据之间具有较高的相似性，而不同类(簇)中的数据之间具有较大的差异性的原则将数据划分成有意义或有用的类(簇)。数据事先不存在类标号。

3. 偏差检测(异常检测)

偏差检测也称异常检测，指通过发现数据集中特殊的变化，寻找孤立点，并对其进行分析，探究原因，以确定是不是事物发生的突变。

4. 关联分析

关联分析指通过挖掘频繁模式来发现大量数据中有趣的关联或相关联系。例如通过购物篮分析，确定哪些商品通常会被一起购买，从而制定交叉销售等营销策略。

5. 预测

预测指用过去和现在的数据去拟合模型，并使用模型预测未来。广义上，预测包含分类，是对类别变量的预测，狭义的预测仅指对连续型变量的预测。

6. 分类

分类指基于已知类别的训练数据构建一个分类模型(分类器)，用于对未知类别的新数据进行分类。所以，用于构建模型的样本数据必须存在类标签。

随着数据类型的多样化、存储技术的进步、计算能力的提升、算法的持续演进、应用需求的驱动，数据挖掘不断突破应用领域，功能也在持续扩展和升级，从最初的结构化数据分析发展到文本挖掘、图像与视频挖掘，并进一步融合多模态数据，以满足日益增长的智能分析和决策需求。

1.3.2　数据挖掘使用技术

数据挖掘的产生和发展一直受应用驱动。随着应用不断拓宽，其所使用的技术也越来越丰富，而且将持续发展，如图 1.10 所示。

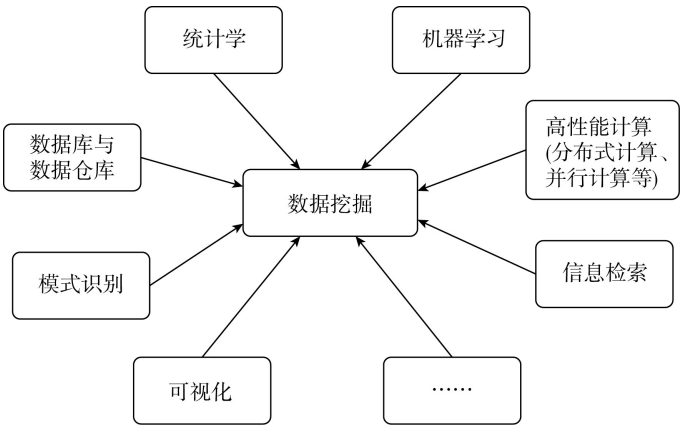


图 1.10　数据挖掘使用技术

从统计学的发展过程看，统计学在自然科学、工业及商业等领域的应用中面临着各种挑战；正是在应对这些挑战的过程中，统计学不断得到充实和发展。随着计算机软硬件技术的飞速发展，数据存储能力无限量地提高，面对海量且形式多样的数据，传统统计学方法在应用时遇到了新的难题。数据挖掘正是统计学适应这一变化的新的发展方向。数据挖掘并不是为了替代传统的统计分析技术，而是统计分析方法的延伸和扩展。Ganesh(2002)认为，从统计学的视角看，数据挖掘可以被看成对大容量复杂数据的计算机自动化的探索和分析，可以被认为是“智能化统计”。因此，统计方法自然成为数据挖掘的一大技术支撑。

传统的统计方法可以分为描述统计和推断统计。描述统计主要对观察到的数据进行汇总、分类和计算，并用表格、图形和指标的形式来反映现象的数量特征。推断统计则以已知的数据(部分的或过去的)去推断未知的数据(整体的或未来的)。这两类方法正好符合数据挖掘两大任务(描述和预测)的需要，数据挖掘把统计学技术与计算机技术相结合，从数据中发现有用的知识。

数据库、数据仓库、大数据分布式存储与高性能计算是数据挖掘的重要基础和支撑技术。它们为数据挖掘提供了数据存储、管理和处理的能力，使得数据挖掘能够从海量数据中提取有价值的信息。相关基础知识将在第 2 章中详细介绍。

1.4 数据挖掘的核心利器：机器学习

机器学习是指计算机利用各种学习算法，从输入的数据中学习，识别复杂的模式，从而做出智能决断。因为学习算法中涉及大量的统计学理论，机器学习与推断统计学的联系尤为密切，所以机器学习有时被称为统计学习理论，尤其是在学术界和理论研究中。

机器学习的基础是数据，核心是各种学习算法，只有通过这些算法，机器才能识别分析这些数据，获得知识，从而不断提升自身性能。因此，机器学习主要研究不同应用场景下应该选用哪种学习算法或研究新的学习算法以适应新的场景需要。

1.4.1 机器学习分类

机器学习的算法很多，根据学习方式不同，可以分为有监督学习(supervised learning)、无监督学习(unsupervised learning)、半监督学习(semi-supervised learning)和强化学习(reinforcement learning)。

1. 有监督学习

用于有监督学习训练的数据集包含输入(特征)和输出(目标)，也称为有标记的数据集。从有标记数据集中根据输入和输出学习得到一个模型，即为有监督学习。当新的数据输入时，可以根据这个模型预测结果。由于训练集中存在目标，因此学习得到的模型可以使用历史数据进行验证，从而起到监督的作用。有监督学习算法主要应用于分类和回归，如决策树、支持向量机、朴素贝叶斯、Logistic 回归、神经网络等。

2. 无监督学习

用于无监督学习训练的数据集只包含输入(特征)，而没有输出(目标)，也称为无标记数据集。在无标记数据集中通过学习进行归纳，获得数据分布特征或数据与数据之间的关系，即为无监督学习。由于训练数据不存在目标，因此学习得到的模型不能使用历史数据进行验证，从而无法监督。无监督学习算法主要应用于聚类、降维和关联分析等，如 K-均值聚类(K-Means Clustering)、层次聚类(Hierarchical Clustering)、主成分分析(PCA，Principal Component Analysis)、Apriori 算法等。

3. 半监督学习

有两个数据集用于半监督学习，一个为有标记的数据集，一个为无标记的数据集，通常无标记数据集的数据量要远远大于有标记数据集的数据量。如上所述，如果单独使用有标记数据集，我们能够生成有监督模型；单独使用无标记数据集，我们能够生成无监督模型。为了最大限度利用现有数据的信息，我们希望使用两个数据集进行学习。用户可以在有标记数据集中加入无标记数据，增强有监督学习的效果，如半监督支持向量机；也可以在无标记数据集中加入有标记数据，增强无监督学习的效果，如半监督聚类。一般而言，半监督学习侧重于在有标记数据集中加入无标记数据来增强学习效果，适用于现实场景中获取标注数据成本高，但未标注数据丰富的情况。

4. 强化学习

强化学习是智能体(agent)在尝试的过程中学习在特定的环境下选择哪种行动可以得到最大的回报。如图 1.11 所示，智能体在学习的过程中选择一个动作，环境接受该动作后状态发生变化，同时产生一个强化信号(奖励或惩罚)，反馈给智能体，智能体根据强化信号和环境当前状态再选择下一个动作，选择的原则是使受到的正强化(奖励)最大。智能体当下选择的动作不仅影响当下的强化值，而且影响环境下一时刻的状态及最终的强化值。

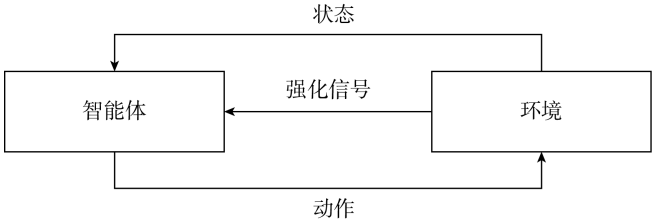


图 1.11　强化学习示意图

1.4.2　机器学习与数据挖掘的关系

数据挖掘旨在从大量数据中发现有价值的信息和模式，机器学习为数据挖掘提供了强大有效的算法和技术来实现这一目标。通过自动学习数据中的规律和模式，机器学习使得数据挖掘过程变得更加高效和智能化，从而在各行各业中得到广泛应用。

机器学习推动了数据挖掘方法和技术的不断演进。

(1) 随着数据量的增长，传统的数据分析方法无法有效处理大量数据。机器学习能够

在大数据环境下高效地发现数据中的模式和关系，应对高维度、海量数据的问题。

(2) 机器学习使得数据挖掘能够实时处理和分析数据，尤其在金融、互联网、智能制造等领域，帮助企业做出动态决策。

(3) 机器学习使得数据挖掘不仅限于结构化数据(如数据库中的表格数据)，还能够处理非结构化数据(如文本、图像、视频等)。例如，自然语言处理(NLP)和计算机视觉的机器学习技术，使得从文本分析到图像识别的应用场景得到极大的拓展。

随着人工智能技术的不断发展，机器学习与数据挖掘的结合将继续推动技术创新。

(1) 深度学习与强化学习：使得数据挖掘能够处理更加复杂的任务，如自动驾驶、智能决策等。

(2) 自监督学习与迁移学习：自监督学习能够减少对大量标注数据的依赖，而迁移学习则允许模型从一个领域迁移到另一个领域，进一步提升数据挖掘的灵活性和应用范围。

(3) 边缘计算与实时数据挖掘：边缘计算结合机器学习将使得数据分析可以在数据源附近进行实时处理，适应物联网等实时数据挖掘的需求。

1.5 数据挖掘应用

数据挖掘从一开始就是面向应用的，随着各行各业信息化的持续发展，数据挖掘应用领域也在不断发展和深化。目前，数据挖掘在金融、电信、零售与电子商务、政府政务、医疗、科学等领域都有应用。

1.5.1 金融领域的数据挖掘

银行、证券和保险等金融领域，信息化建设较早，积累了大量的数据，是数据挖掘的重要应用领域，典型的应用有：金融风险分析、金融产品交叉销售、客户管理分析、洗黑钱等金融犯罪识别等。金融交易活动过程很可能存在洗黑钱等犯罪行为，把可能与侦破有关的数据集成(如金融机构交易数据库、犯罪历史数据库等)，运用合适的数据挖掘方法(数据可视化、孤立点分析等)，检测异常模式，可以为犯罪行为识别提供快速准确的参考。

银行业利用数据挖掘技术最集中的两个方面是风险管理和客户管理。风险管理，如信用风险评估，银行可通过建立信用风险模型，评估贷款申请人或信用卡申请人的风险，根据评估结果来决定是否接受申请，并确定贷款额度或信用额度。客户管理体现在客户生命周期的各个阶段，包括客户获取阶段的客户画像，客户保留阶段的客户细分、客户价值分析及客户流失分析等。在客户保留阶段，根据银行大量的客户基本属性数据、客户存款、贷款、金融产品使用等数据，利用聚类的方法，实现客户细分，将客户有效地划分为不同的类，从而针对每一类客户的特征设计出相应的产品组合、服务模式，以提高客户忠诚度。

证券业利用数据挖掘技术最集中的两个方面是客户管理和量化交易。证券公司可以利用客户个人基本信息、客户交易操作行为数据、软件使用习惯、自选股、常用分析指标等

对客户的理财需求进行挖掘，实现精准营销。量化交易可借助数据挖掘方法，对证券期货市场的海量数据进行分析和挖掘，获得证券期货产品的价格变化规律，得到能带来超额收益的交易策略模型，然后通过分析结果来指导投资，以获得可持续、稳定且高于平均的超额回报。

保险业利用关联挖掘或各种推荐算法可以发现客户购买保险产品的关联与偏好，从而实现交叉销售。保险公司标的受损时，通过挖掘已有标的定损数据，可以对现有标的损失进行精确估计和预测，从而实现保险智能定损。随着保险业的发展，保险欺诈问题也日益突出，给保险公司和社会带来了极大危害。利用数据挖掘方法，分析并识别欺诈行为的特征，可以对保险欺诈行为进行实时监测与预警，从而促进保险业健康有序发展。

1.5.2　电信领域的数据挖掘

随着信息技术的迅速发展，电信业从 4G 时代进入 5G 时代，在电信业务迅速发展的同时，电信行业的竞争也日益激烈。面对国内、国际电信业激烈的竞争态势，各大电信运营商纷纷使用数据挖掘技术了解行业动向、分析业务模式、洞察客户需求，实现精细化的管理和精准营销，提升自身服务质量，从而提高客户的满意度和忠诚度，增强竞争优势。

在客户关系管理方面，运营商使用数据挖掘可以对客户进行画像以提供个性化的业务推荐，可以对客户进行细分以发现不同价值的客户群体特征，可以通过客户流失分析制订相应的挽留策略，可以对客户之间的社会关系进行社交网络分析以获取潜在客户和保持现有客户，可以对客户流量使用进行异常识别，挖掘导致其流量异常的恶意程序和恶意 App，以减少用户不必要的损失，并防止其他用户遭受同样的恶意攻击。在市场营销方面，可以使用关联挖掘进行电信业务的交叉销售。

运营商对网络信令数据进行挖掘，可以预测网络流量峰值，预警异常流量，防止网络堵塞和宕机，从而提高网络服务质量，提升用户体验。对移动用户的位置信息进行挖掘，与相关企业合作，可以提供基于位置的相关服务，如餐饮推荐、优惠券推送，这将改变运营商的盈利模式，而且具有非常广阔的应用前景。

1.5.3　零售与电子商务领域的数据挖掘

零售业的发展经历了从百货商店到超级市场、连锁商店、电子商务，再到如今线上线下相结合的“新零售”，积累了大量关于采购、销售、客户、物流等方面的数据。数据挖掘在零售与电子商务领域的应用非常广泛，包括用户行为分析、个性化推荐、产品分析、广告追踪与优化、精准营销等，如顾客去商场购物的场景中，商场基于移动手机与 Wi-Fi 结合的数据，根据顾客所有的行动轨迹，分析顾客光顾的时间和频率、行径路线、驻留时间和地点，实现精准营销。

随着新零售业态的发展，线上线下系统对接和数据融合，零售企业借助数据挖掘技术可以对消费者全过程数据进行描述和产业链营销重构，实现数据化运营，探索新商业模式，建立新市场增长点。

1.5.4 政府政务领域的数据挖掘

政府信息化经过多年建设，已经有效实现了信息化办公。从 2015 年国家发布《促进大数据发展行动纲要》(国发〔2015〕50 号)开始，我国政府已将政务信息系统整合及共享提升到国家战略层面，对互联网+政务服务体系的建设给出了明确指导意见和时间点要求。

国防、教育、公安、民政、司法、财政、交通运输、农业、商务、文化和旅游等政务部门信息系统的整合与共享，使数据挖掘的应用更加广泛。结合数据挖掘技术，政府加强统筹规划，实现智慧交通、智慧安防、智慧旅游等，加强智慧城市建设，使政务工作更高效、更开放、更透明。

1.5.5 医疗领域的数据挖掘

医疗领域积累了大量数据，尤其是海量的非格式化数据。数据挖掘在医疗领域的应用，主要集中在药品研发、疾病治疗、公共卫生管理、居民健康管理和健康影响因素分析等方面。

在药品研发方面，医药公司可以借助数据挖掘，在研发初期通过建模确定最有效率的投入产出比，配备最佳资源；在药物临床试验阶段，及时预测临床结果，选择最优药物。在疾病治疗方面，医生可以结合病人体征数据、费用数据和疗效数据进行挖掘，以确定在临床上对病人最有效和最具有成本效益的治疗方案。而且，对于医疗影像数据的分析和挖掘，会极大减少医生的工作量，提高医疗效率。

在公共卫生管理、居民健康管理方面，卫生部门基于覆盖全国的电子病例数据进行挖掘，可以快速检测传染病，有效监测疫情，并提供有针对性的公众健康咨询，提高公众健康风险意识，降低传染病感染风险。

1.5.6 科学领域的数据挖掘

天文学、气象学、地质学、生物学等各领域使用全球定位系统、卫星遥感器及新一代生物学数据采集技术，收集了海量的包含时间和空间信息的高维数据、流数据和异构数据。早期，数据挖掘应用于天文学，在短短 4 小时内发现的行星超过 20 多位天文学家 4 年的研究成果。

人类拥有 23 对染色体，约含有 30 亿对 DNA 碱基。1975 年，英国科学家 Frederick Sanger 发明了 Sanger 测序技术，由此开启了基因测序的新篇章。1990 年，由全球多个国家共同参与的人类基因组计划正式启动，被称为人类三大科学计划之一，旨在为这 30 亿对碱基构成的人类基因测序。数据挖掘技术应用于基因测序后，极大降低了测序成本，提升了测序速度。得益于此，从疾病的筛查、诊断到治疗，越来越多的临床基因检测项目落地，如新生儿疾病筛查、遗传病筛查、肿瘤易感基因筛查和肿瘤个性化用药等。

1.6　练习与拓展

即测即评

扫右侧二维码，完成客观题自测题。

即测即评

练习

1. 什么是数据挖掘？请结合实例加以说明。
2. 简述第四代数据挖掘系统的特点。
3. 什么是云计算？
4. 什么是大数据？如何理解大数据被认为是下一个社会发展阶段的石油和金矿。
5. 什么是多模态数据挖掘？
6. 请分析说明 Fayyad 过程模型。
7. 请分析说明 CRISP-DM 过程模型。
8. 数据挖掘的功能有哪些？
9. 结合数据挖掘使用技术，请分析其与相关学科之间的关系。
10. 什么是机器学习？按学习方式不同，机器学习可以分成哪几种？分别具有什么特点？
11. 请举例说明教材中提到的数据挖掘应用领域，并谈谈你的理解。
12. 除了教材中提到的数据挖掘应用领域，请思考还有哪些应用领域，并举例说明。

拓展

1. 检索近几年数据挖掘国际学术会议的入选论文，分析数据挖掘研究现状及热点问题。
2. 查阅相关资料，了解多模态数据挖掘涉及的核心技术。
3. 查阅相关资料，了解半监督学习的常用方法。
4. 结合 CRISP-DM 过程模型，自选一个感兴趣的商业问题，以小组为单位，制订一份数据挖掘项目计划。