第1章
大数据分析与挖掘概论

随着 Web 2.0、物联网等信息技术的发展，组织内部积累了各种业务系统中形成的结构化数据，同时也能够获取海量的非结构化数据，如用户的评论数据、产品和设备状态的实时数据、各类社交媒体平台的数据。这些非结构化数据呈现出海量、多样、高速等特点，需要使用以机器学习为代表的数据挖掘技术进行大数据分析与处理，为组织的智能化决策提供帮助。

1.1 大数据及其应用

1.1.1 大数据的概念

我们正生活在大数据时代，互联网时刻在产生来自商业、社会、科学和工程、医学以及日常生活的方方面面的大数据。数据的爆炸式增长是技术驱动数字社会转型和功能强大的数据存储工具快速发展的结果。全世界范围内的商业活动产生了海量的结构化数据，如产品销售、金融交易、电子商务、组织运营以及顾客服务等业务产生的数据。科学和工程实践持续地从遥感、过程测量、科学实验、系统实施、工程观测和环境监测中产生海量的实时数据。医疗领域的医疗记录、病人监护和医学图像也是重要的健康大数据来源。社会化媒体发展，使各类商务平台、政务平台、网络社区、搜索引擎和社交软件等已经成为日趋重要的非结构化数据来源，包括文本、图片、视频、Web 页等，蕴含着重要的价值。

大数据的概念起源于 2008 年 9 月 *Nature* 杂志刊登的名为“Big Data”的专题。2011 年，*Science* 杂志也推出专刊 *Dealing with Data* 对大数据计算问题进行讨论。维克托·迈尔-舍恩伯格(Viktor Mayer-Schönberger)及肯尼思·库克耶(Kenneth Cukier)所著的《大数据时代》一书中，提出大数据是摒弃了抽样调查而采用所有数据进行分析处理的方法。Gartner 将“大数据”定义为需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

目前，大数据并没有一个明确的定义。从狭义层面，大数据通常被理解为“用现有的一般技术难以管理的大量数据的集合”。然而，该定义仅着眼于“大数据”一词的数据相关性质，并不能全面解释大数据相关的问题和内容。从广义层面，所谓“大数据”，包括因具备 4V 特征而难以进行管理的数据，也包括对这些数据进行存储、处理、分析的技术，以及能够通过分析这些数据获得实用价值的人员、组织和系统。其中，“存储、处理、分析的技术”指的是用

于大规模数据分布式处理的框架 Hadoop、具备良好扩展性的 NoSQL(非关系型数据库),以及机器学习(machine learning)和统计分析方法等。“能够通过分析这些数据获得实用价值的人员、组织和系统”指的是能够对大数据进行有效存储和运用的技术人员、数据分析公司和管理信息系统。

1.1.2 大数据的特征

大数据具有四个维度的特征,包括规模性(volume)、多样性(variety)、高速性(velocity)和价值性(value),简称 4V 特征。

(1) 规模性。从大数据的定义可知,规模性体现在数据的存储和计算均需要耗费海量的计算资源,现在来看,基本上是指从几十 TB 到几 PB 的数量级。随着数据处理技术的进步,这个数值也在不断变化。若干年后,也许只有几个 EB 数量级的数据量才称得上是大数据。中商产业研究院发布的《2023 年全球及中国数据产量预测分析》数据显示,全球数据产量由 2019 年的 42 ZB 增长至 2022 年的 81.3 ZB,复合年均增长率达 24.6%。另外,对于不同的应用领域,大数据的数据量也有所不同。例如互联网大数据要比传统制造业大数据量大得多,而随着智能制造中大量使用物联网、移动计算等“5G+工业互联网”技术,现代制造业中的大数据同样具有海量性。

(2) 多样性。多样性指的是大数据来源和形式的多样性,体现为多模态数据。大数据类型十分丰富,主要包括结构化数据和非结构化数据。随着 Web 2.0 技术、移动互联网的发展,网络中产生了大量非结构化数据,如位置信息、文本、图片、音频和视频等。现代企业运营中会产生生产、销售、库存、财务、人事等结构化数据,还会收集和使用海量非结构化数据,如网站日志、社交网络、全球定位系统(GPS)位置数据和温湿度等传感器数据,以及图片、语音和视频等各种非结构化数据。这些类型繁多的多源异构数据,对大数据分析处理提出了许多挑战。

(3) 高速性。高速性是指大数据产生和更新的频率很快。例如,POS 机(销售点终端机)交易数据、电商网站点击流、购买数据和评论数据、社交网站中用户发布的推文数据、搜索引擎的实时搜索数据、移动应用商店的 App 下载数据、市场遍布全球的传感器和摄像头所采集的数据等,这些数据以极高的速度产生、存储和利用,对这些数据的分析和处理颇具挑战。大数据分析处理的需求推动了流数据处理等新技术的发展,产生实时分析结果,用于指导生产和生活实践。

(4) 价值性。大数据蕴含着重要价值,价值性体现大数据分析与应用的目的。通过深入的大数据分析与挖掘,可以为企业与政府等组织的经营和管理决策提供有效支持,创造巨大的经济社会价值。同时,大数据也具有价值密度低的特点。一般价值高低与数据总量的大小成反比,以视频为例,一部 1 小时的视频,在不间断的监控中,有用数据可能仅有几秒。如何通过强大的人工智能和机器学习算法实现大数据的价值提取,成为目前大数据分析亟待解决的难题。

1.1.3 大数据的应用

新兴信息技术发展及在各个行业的应用,使得行业大数据的应用越来越深入,产生了重

要的经济社会价值。如金融大数据、消费大数据、工业大数据、城市大数据、政府治理大数据等应用领域，实现数据驱动的管理决策，显著提升企业管理和政府治理水平与能力。

(1) 金融大数据。金融领域是数字化转型较为成熟的领域，线上线下业务积累了海量的用户数据和业务数据。如中国建设银行加速推进业务、数据、技术“三大中台”建设，数据中台方面，以共享数据资源和能力为核心，夯实多源异构数据的统一数据基础，持续丰富智能数据产品货架，打造全域数据视图。中国建设银行网站显示，截至 2022 年底，建设银行金融科技投入 232.90 亿元，占营业收入的 2.83%；线上用户数超过了 5 亿户，其中手机银行用户数达到 4.4 亿户，月均月活数 1.32 亿户；“建行生活”客户数达 1 亿规模。大数据在金融行业的应用包括精准营销、风险管控、决策支持、效率提升和产品设计等方面。

(2) 消费大数据。电子商务日益成为人们消费的主要渠道。电商数据较为集中，数据量足够大，数据种类较多，使用电商数据挖掘消费者需求以及高效整合供应链满足其需求的能力越来越重要。消费大数据的应用包括预测流行趋势、消费趋势、地域消费特点、客户消费习惯、消费热点、影响消费的因素等。

(3) 工业大数据。工业互联网的主要特征是智能和互联，通过充分利用信息通信技术，把产品、机器、资源和人有机结合在一起，推动制造业向基于大数据分析与应用的智能化转型。随着智能制造时代的到来，工业大数据的应用将成为提升制造业生产力、竞争力、创新能力的关键要素。工业大数据的应用包括提升工厂运营效率、优化供应链、创新商业模式、提升产品质量等方面。

(4) 城市大数据。城市大数据是指在城市发展过程中形成的数据资源，包括人口、交通、环境、经济、社会、文化等方面的信息。对这些数据的收集、整理、分析与应用，可以为政府决策提供科学依据，为社会公众提供服务，促进城市可持续发展，显著提升智慧城市治理水平。以交通大数据为例，一方面可以利用交通传感器数据了解车辆通行密度，合理进行道路规划；另一方面可以利用交通大数据来实现即时交通信号调度，提高线路运行能力。

(5) 政府治理大数据。政府利用经济、资源、公众等方面大数据，可以显著提升政府智慧治理能力，优化资源配置。如政府依据经济发展大数据，可以了解地区经济发展情况、产业发展情况、人民生活状况等，科学地制定宏观政策。政府使用大数据舆情监控，以减少群体性事件，提升社会治理水平。

1.2　大数据分析与挖掘的相关概念

1.2.1　大数据分析与挖掘的重要性

1. 商业实践与现实需求

数字技术驱动的商业实践积累了海量的数据，数据的广泛性和多样性产生了“数据丰富，但信息贫乏”的现象。快速增长的海量数据被收集、存放在数据库中，没有强有力的数据分析工具，理解它们已经远远超出了人的能力。这样，商业决策通常不是基于数据库中含有丰富信息的数据，而是基于决策者的直觉，原因在于决策者缺乏从海量数据中提取有价值知

识的工具。尽管学术界和产业界在开发专家系统与知识库系统方面已经做出很大的努力，但是这种系统通常依赖用户或领域专家人工地将知识输入知识库，这一过程常常有偏差和错误，并且费用高、耗费时间。

大量数据不仅仅累积在数据库和数据仓库中，20 世纪 90 年代，万维网和基于 Web 的数据库（如 XML 数据库）开始出现。诸如万维网和各种互联的、异种数据库等基于互联网的全球信息库已经出现，并在信息产业中扮演极其重要的角色。数据和信息之间的鸿沟越来越宽，通过集成信息检索、数据挖掘和信息网络分析技术来有效地分析这些不同形式的数据成为一项具有挑战性的任务。这就要求系统地开发大数据挖掘（big data mining）技术，对海量的结构化大数据和非结构化大数据进行高效处理，以实现数据驱动的商业决策。

2. 解决方案

数据的爆炸式增长、广泛可用和巨大数量使我们的时代成为真正的数据时代。在这个时代，我们急需功能强大和通用的工具，以便从这些海量数据中发现有价值的信息，把它们转化成有组织的知识。大数据分析与挖掘的手段通常有两种：一是通过数据仓库，进行联机分析处理（OLAP）；二是通过数据挖掘建模从大规模数据中抽取有用的知识。OLAP 是一种分析技术，具有汇总、合并和聚集以及从不同的角度观察信息的能力。尽管 OLAP 工具支持多维分析和决策，但是对于深层次的分析，仍然需要其他分析工具，如提供数据分类（classification）、聚类（clustering）、离群点/异常检测和刻画数据随时间变化等特征的大数据挖掘技术。本书将重点介绍大数据挖掘的建模方法。

数据挖掘可以看作信息技术自然进化的结果。数据库和数据管理在一些关键功能的开发上不断发展：数据收集和数据库创建、数据管理（包括数据存储和检索）和高级数据分析（包括数据仓库和数据挖掘）。作为高级数据分析的手段，数据挖掘可以把大型数据集转换成有价值的知识。像 Google（谷歌）搜索引擎每天接受数亿次查询，每个查询都被看作一个事务，用户通过事务描述他们的信息需求。随着时间的推移，搜索引擎可以从这些大量的搜索查询中学到新颖的、有用的知识。如 Google 的 Flu Trends（流感趋势）使用特殊的搜索项作为流感活动的指示器，发现了搜索流感相关信息的人数与实际具有流感症状的人数之间的紧密联系。使用聚集的搜索数据，Google 的 Flu Trends 可以比传统的系统早两周对流感活动作出评估。因此，大数据挖掘已经成为大数据时代商业决策的最重要技术和工具。

1.2.2 相关概念

1. 大数据挖掘

大数据挖掘是从海量的结构化数据和非结构化数据中挖掘出隐含的、未知的、用户可能感兴趣的、对决策有价值的知识和规则的过程。大数据挖掘涉及机器学习、统计学、数据库与数据仓库、信息检索、算法、模式识别、分布式计算、可视化技术等核心技术；数据源包括数据库、数据仓库、Web 数据、文本、多媒体数据、空间数据、时序数据等结构化数据和非结构化数据。

大数据挖掘是人工智能和数据库领域研究的热点问题，其任务有关联分析、聚类分析、分类分析、离群点分析等。大数据挖掘的应用领域包括商务管理、生产控制、市场分析、运营管理、政府治理、智慧城市管理、工程设计和科学研究等方面。

2. 知识发现

知识发现(knowledge discovery in database，KDD)是应用特定的数据挖掘算法按指定方式和阈值抽取有价值的知识，以及评价解释模式的一个循环反复过程。知识发现的过程包括数据获取、数据预处理、数据挖掘、结果评价与解释等步骤，从这一视角来看，数据挖掘是知识发现的一个环节。但是现实中，往往将两者等同起来，数据挖掘多为统计学、数据分析及管理信息系统领域采用，而知识发现通常用于人工智能、机器学习领域。

3. 商务智能

商务智能(business intelligence)是一个从大规模数据中发现潜在的、新颖的、有用的知识的过程，旨在支持组织的业务运作和管理决策。

商务智能是一套完整的解决方案，它将数据仓库、联机分析处理和数据挖掘等结合起来应用到商业活动中，从不同的数据源收集数据，对数据进行抽取、转换和装载，将所得到的信息存入数据仓库或数据集市，然后使用合适的查询与分析工具、联机分析处理工具和数据挖掘工具对信息进行处理，将信息转变为辅助决策的知识，最后将知识呈现在用户面前，以实现技术服务与决策的目的。

因此，数据挖掘侧重从海量数据中发现隐含的、未知的并有潜在价值的信息，是商务智能最重要的技术基础。商务智能是数据挖掘的应用，目的是为企业提供数据驱动的决策支持。数据挖掘与商务智能的关系如图 1-1 所示。

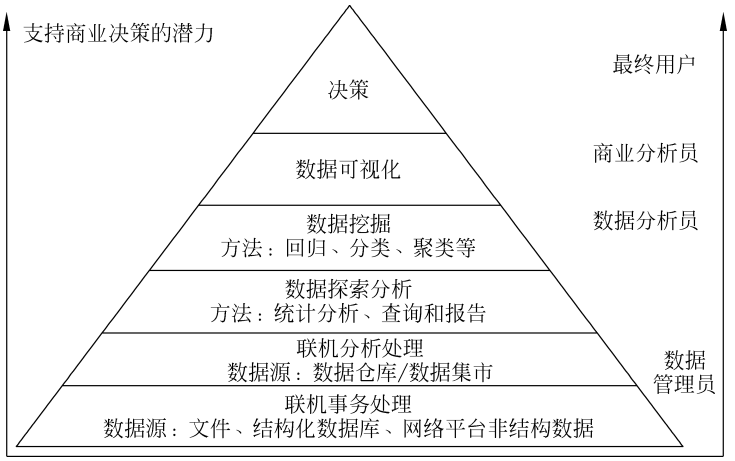

图 1-1　数据挖掘与商务智能的关系

作为商务智能的核心技术，数据挖掘的商业应用体系包括行业应用层、商业逻辑层、数据挖掘算法层(图 1-2)。其中，行业应用层是不同数据挖掘算法所解决的各类商业问题。

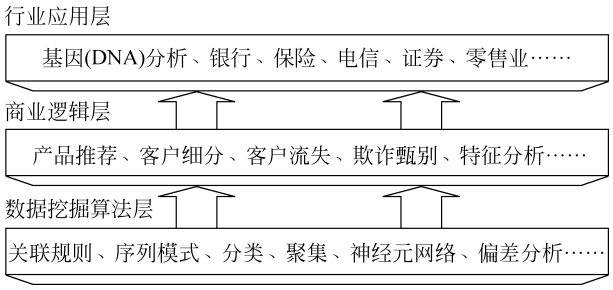

图 1-2　数据挖掘的商业应用体系

1.2.3　大数据分析与挖掘的过程

大数据分析与挖掘的过程包括数据准备、模型构建与评估、结果表达和解释三个环节，具体由七个步骤构成，如图 1-3 所示。

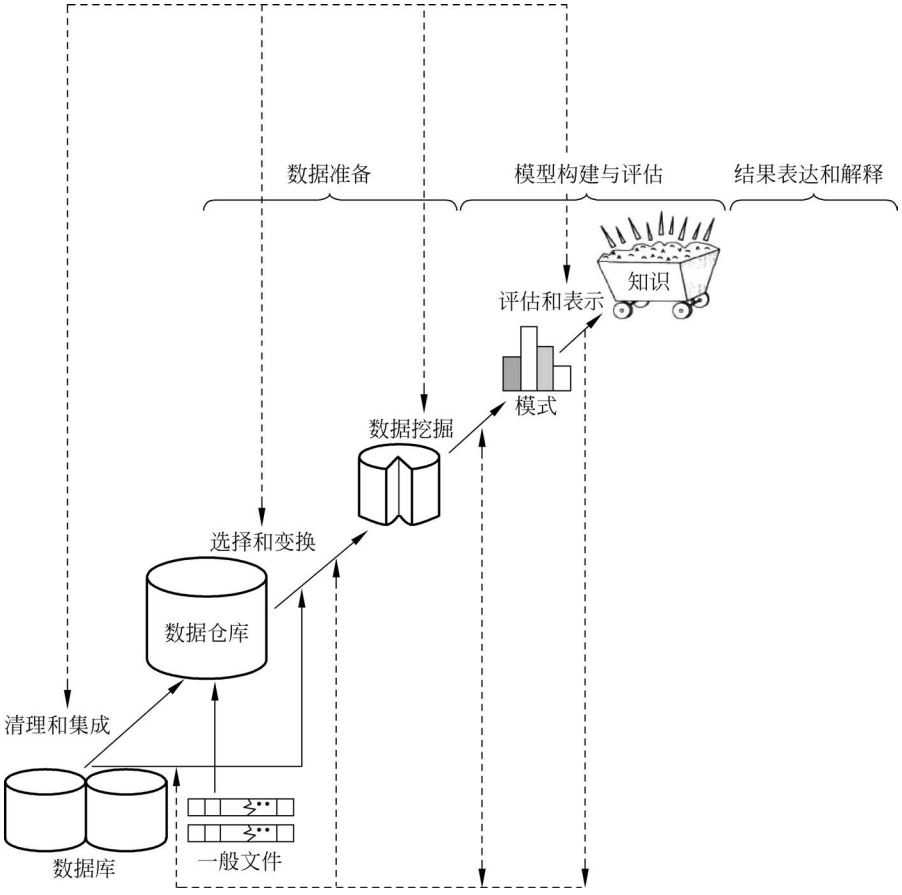

图 1-3　大数据分析与挖掘的过程

(1) 数据清理[消除噪声(noise)和删除不一致数据]。

(2) 数据集成(将多种数据源组合在一起)。

(3) 数据选择(从数据库中提取与分析任务相关的数据)。

(4) 数据变换(通过汇总或聚集操作,把数据变换和统一成适合挖掘的形式)。

(5) 模型构建(使用机器学习、统计学等方法提取数据模式。建立模型是一个反复的过程,需要仔细考察不同的模型以判断哪个模型对商业问题最有用。先用一部分数据建立模型,然后再用剩下的数据来测试和验证这个得到的模型)。

(6) 模式评估(根据某种兴趣度度量,识别真正有趣的模式)。

(7) 知识表示(使用可视化和知识表示技术,向用户提供挖掘的知识)。

步骤(1)～(4)是数据预处理的过程,为大数据挖掘准备数据。模型构建与模式评估阶段可能与用户或知识库交互,然后将有趣的模式提供给用户,或作为新的知识存放在知识库中。有时数据变换在数据选择过程之前进行,特别是在数据仓库的情况下。可能还需要进行数据归约,以得到原始数据的较小表示,而不牺牲完整性。

1.3 大数据分析与挖掘的模式

1.3.1 模式类型

大数据分析与挖掘的任务是发现数据背后的规律和模式。一般而言,这些任务可以分为两类:描述型(descriptive)和预测型(predictive)。描述型挖掘任务是刻画目标数据中数据的一般性质。预测型挖掘任务是对当前数据进行归纳,以便作出预测。大数据分析与挖掘的具体算法和工具类型如图 1-4 所示。

1.3.2 关联规则

关联规则(association rule)挖掘是在大量数据中挖掘数据项之间的关联关系,寻找可靠的频繁模式,其典型的应用就是购物篮分析。

频繁模式是在数据中频繁出现的模式,存在多种类型的频繁模式,包括频繁项集、频繁子序列和频繁子结构。频繁项集一般是指频繁地在事务数据集中一起出现的项的集合,如超市中被许多顾客频繁地一起购买的商品组合(如牛奶和面包、啤酒和尿布、相机和存储卡等)。频繁子序列是频繁出现的项的序列模式,如顾客倾向于先购买便携计算机,再购买数码相机,然后购买内存卡这样的模式就是一个频繁序列模式。子结构可能涉及不同的结构形式(例如,图、树或格),可以与项集

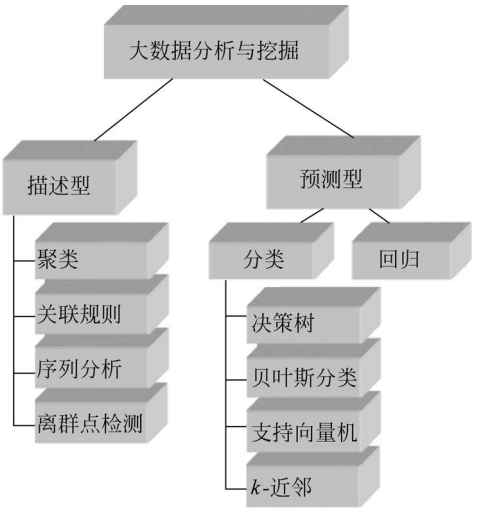

图 1-4　大数据分析与挖掘的具体算法和工具类型

或子序列结合在一起。如果一个子结构频繁地出现，则称它为频繁结构模式。挖掘频繁模式以发现数据中有趣的关联关系。

例：关联分析。假设作为超市经理，你想知道哪些商品经常被一起购买（即在相同的事务中）。从超市的事务数据库中挖掘出来的这种关联规则的一个例子是

buys(X，“computer”)=>buys(X，“software”)[support=1%，confidence=50%]

其中，X 代表顾客。50%的置信度意味着如果一位顾客购买计算机，则购买软件的可能性是 50%。1%的支持度意味着所分析的所有事务的 1%显示计算机与软件一起被购买。这个关联规则涉及单个的属性或谓词（即 buys）。通常，一个关联规则如果同时满足最小支持度阈值和最小置信度阈值，则认为是一个有趣的关联关系。

关联规则挖掘在很多领域有广泛应用，如产品推荐、网络入侵检测、基因分析、医疗诊断等。

1.3.3 分类

分类是一种典型的有监督学习（supervised learning）问题，用于建立数据特征和数据类别之间映射关系的模型，以便使用模型预测类标号未知的对象的类标号。分类的过程一般包括分类器训练和预测两个阶段。通常会将已有数据集划分为训练集和测试集两个部分，训练集用来训练分类器，测试集用来评估分类器的效果，训练集中的每一个样本除了包含一些特征外，还有一个标注好的标签类别。分类器训练完成后，能够对没有类别标签的样本进行预测，得到合适的标签。分类器的构建流程如图 1-5 所示。

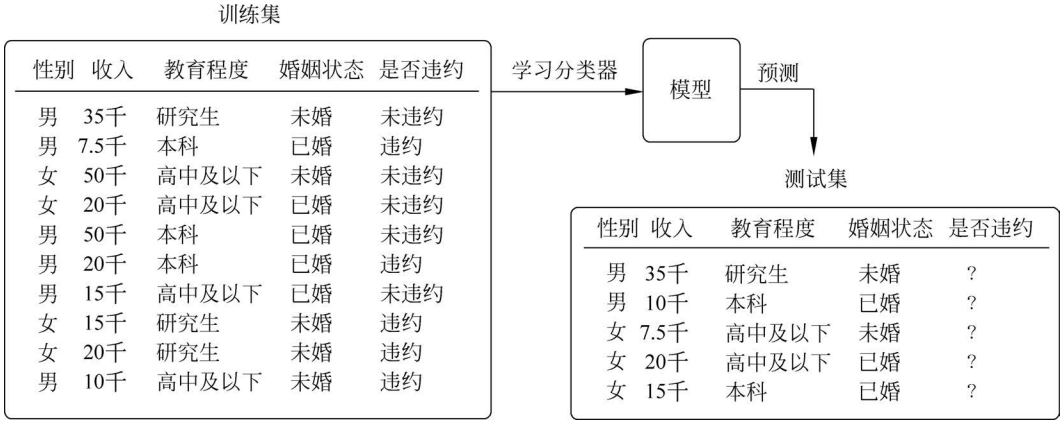

图 1-5 分类器的构建流程

分类模型等价于寻找一个函数 f，不同分类模型体现在对 f 形式的假设不同，可以用多种形式表示，如逻辑回归（logistic regression）、决策树、朴素贝叶斯（Naïve Bayesian）、支持向量机和神经网络等。图 1-6 是决策树与神经网络。决策树是一种类似于流程图的树结构，其中每个结点代表在一个属性值上的测试，每个分枝代表测试的一个结果，而树叶代表类标签。当用于分类时，神经网络是一组类似于神经元的处理单元，单元之间加权连接。

分类模型广泛应用于疾病预测、信用风险评估、产品推荐、垃圾邮件检测等场景中。

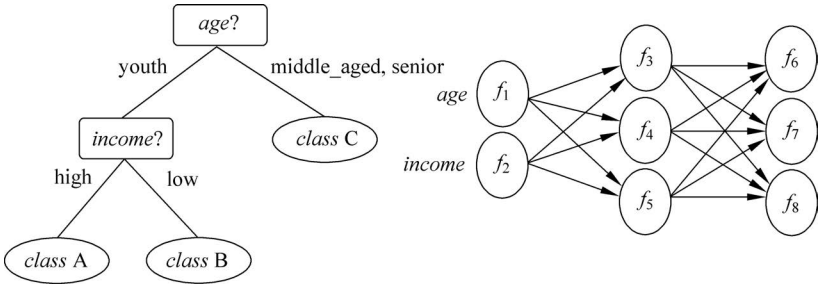

图 1-6　决策树与神经网络

1.3.4　回归

回归是一种确定两种或两种以上变量间相互依赖的定量关系的有监督学习方法，用于建立连续值函数模型。回归模型用来预测缺失或难以获得的数据值，而不是（离散的）类标号。在一个回归模型中，需要关注或预测的变量叫作因变量（响应变量或结果变量），用来解释因变量变化的变量叫作自变量（解释变量或预测变量）。

回归分析（regression analysis）按照涉及的变量的多少，可分为一元回归分析和多元回归分析；按照因变量的多少，可分为简单回归分析和多重回归分析；按照自变量和因变量之间的关系类型，可分为线性回归分析和非线性回归分析。

如图 1-7 所示，假设根据产品质量数据预测产品的用户满意度，这是一个典型回归分析的例子，因为所构造的回归模型将预测一个连续型的函数值。

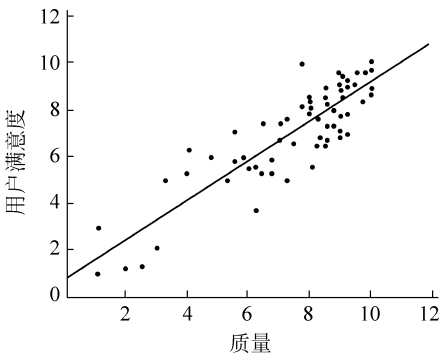

图 1-7　回归分析

回归模型广泛应用于收入预测、销量预测、库存预测和绩效预测等类别为连续值的场景中。

1.3.5　聚类

聚类是对数据集中相似的样本进行分组的过程，是一种典型的无监督学习（unsupervised learning）方法。聚类中每个组称为一个“簇”，每个簇的样本对应一个潜在的类别。聚类分析是对未知类别标签的数据进行直接处理，其目标是使簇内样本的相似性最高，簇间样本的相似性最低。每一个簇看成一个类别，可以简化数据，从中寻找数据的内部结构。常见的聚类算法有 K-means（K-均值）、层次聚类和密度聚类等，如图 1-8 所示。

聚类分析广泛应用于商业、金融、医疗、教育、电商、旅游等行业涉及的市场细分、客户分类、产品定位、用户画像、信用评级等方面。

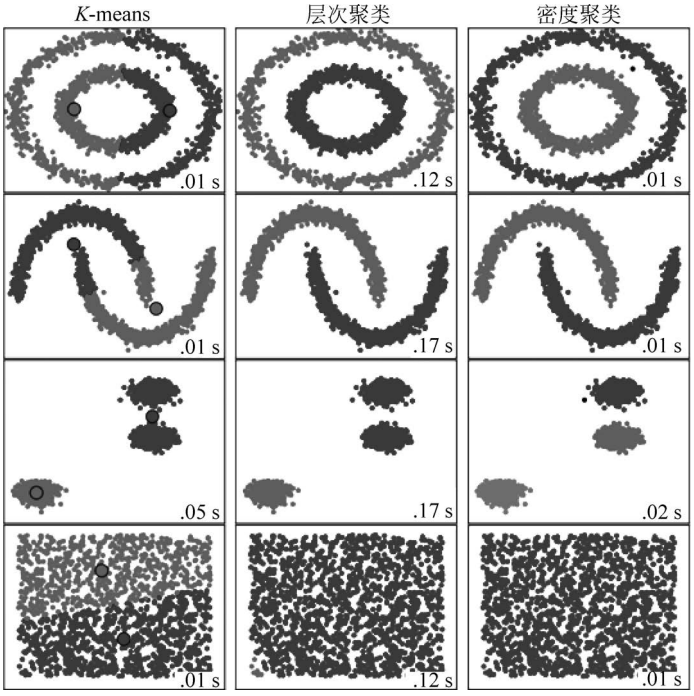

图 1-8　常见的聚类算法

1.4　大数据分析与挖掘技术

数据挖掘吸纳了诸如统计学、机器学习、模式识别、数据库和数据仓库、信息检索、可视化、算法、高性能计算和许多应用领域的大量技术，如图 1-9 所示。

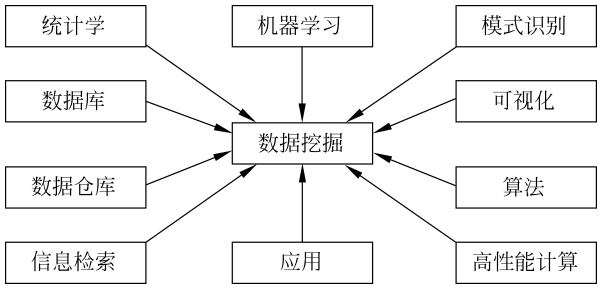

图 1-9　数据挖掘的相关技术

1.4.1　机器学习

机器学习是数据挖掘的核心技术，主要研究计算机程序基于数据自动地学习识别复杂的模式，并作出智能的决断。机器学习可用来找到将特征 X 和 Y 关联的模型 F，从数据到特征 X 的步骤通常是人工完成的(特征工程)。

机器学习的任务主要分为以下三类。

1. 有监督学习

有监督学习要求数据集中的样本带有一个输出标签，模型的目标是找到一个样本到标签的最佳映射，典型的有监督学习包括回归和分类。前者的标签是连续型的，如线性回归、岭回归、LASSO(Least Absolute Shrinkage and Selection Operator，最小绝对值收敛和选择算子算法)回归等算法；后者的标签是离散型的，如决策树、朴素贝叶斯、神经网络、支持向量机、Boosting/Bagging 等算法。深度学习(deep learning)是机器学习领域的一个新的研究方向，基于神经网络算法的扩展产生卷积神经网络(convolutional neural network，CNN 或 ConvNet)、递归神经网络(recursive neural network，RecNN)等算法。深度学习的应用非常广泛，包括图像识别、语音识别、自然语言处理、推荐系统等领域。

2. 无监督学习

无监督学习不要求数据集中的样本带有标签，它根据用户的兴趣来刻画数据的某种统计规律。典型的无监督学习包括聚类和关联规则挖掘等。聚类算法包括划分方法、层次聚类、密度聚类和谱聚类等；关联规则挖掘算法包括 Apriori 算法、FP 树(Frequent Pattern Tree，频繁模式树)等。

3. 强化学习

强化学习(reinforcement learning，RL)是一种特殊的机器学习方法，它让用户在学习过程中扮演主动角色，用于描述和解决智能体在与环境的交互过程中通过学习策略以达成回报最大化或实现特定目标的问题。强化学习可分为基于模式的强化学习(model-based RL)和无模式强化学习(model-free RL)，以及主动强化学习(active RL)和被动强化学习(passive RL)。其中，主动强化学习方法可能要求用户(例如领域专家)对一个来自未标记的实例集或由学习程序合成的实例进行标记。给定可以要求标记的实例数量的约束，目的是通过主动地从用户获取知识来提高模型质量。

1.4.2　统计学

统计学研究数据的收集、分析、解释和表示。大数据挖掘与统计学具有紧密联系。统计模型是一组数学函数，它们用随机变量及其概率分布刻画目标对象的行为。统计模型广泛用于数据和数据类建模，如在数据特征化和分类这样的大数据挖掘任务中，可以建立目标类的统计模型。数据挖掘任务也可以建立在统计模型之上，如使用统计模型对噪声和缺失的数据值建模，大数据在集中挖掘模式时，可以使用该模型来帮助其识别数据中的噪声和缺失的数据值。

统计学研究开发许多使用数据和统计模型进行预测的工具。统计学方法可以用来汇总或描述数据集，帮助从数据中挖掘各种模式，以及理解产生和影响这些模式的潜在机制。推理统计学(或预测统计学)用某种方式对数据建模，解释观测中的随机性和确定性，并用来提取关于所考察的过程或总体的结论。

统计学方法也可以用来验证数据挖掘结果，如建立分类或预测模型之后，应该使用统计假

设检验来验证模型。统计假设检验使用实验数据进行统计推断，如果结果不大可能随机出现，则为统计显著的。如果分类或预测模型有效，则该模型的描述统计量将增强模型的可靠性。

1.4.3 数据库与数据仓库

数据库系统研究关注最终用户创建、维护和使用数据库。数据库系统研究者已经建立了数据建模、查询语言、查询处理与优化方法、数据存储以及索引和存取方法等数据库的功能及工具。用户通过查询语言、用户界面、查询处理优化和事务管理，可以方便、灵活地访问数据。数据库系统因其在处理非常大的、相对结构化的数据集方面的高度可伸缩性而闻名。政府、企业等组织运营数据大多数是采用数据库进行管理，因此数据库也称为大数据挖掘重要的数据源和数据管理工具。

数据仓库是一种将多个异构数据源在单个站点以统一的模式组织存储的数据存储结构，让用户能运行查询、产生报告、执行分析，以支持管理决策。数据仓库的奠基人威廉·H.英蒙(William H. Inmon)对数据仓库的定义是：数据仓库是支持管理决策过程的、面向主题的、集成的、随时间变化的、用来支持管理人员决策的数据集合。数据仓库的底层是多个数据源，一般情况下，这些数据源可以是关系数据或其他类型数据，如平面文件(flat files)、XML(可扩展标记语言)文档等，从数据源中按照统一的规则抽取数据，经过数据清理、数据抽取和转换、数据过滤、数据汇总等过程，将数据转换成数据仓库所需的形式，并将其加载到数据仓库。数据仓库是为联机分析处理、大数据挖掘提供海量数据存储、数据组织的容器和解决数据集成问题的关键技术。

许多大数据挖掘任务都需要处理大型数据集，甚至是处理实时的快速流数据。因此，大数据挖掘可以很好地利用可伸缩的数据库技术，以便获得在大型数据集上的高效率和可伸缩性。此外，大数据挖掘任务也可以用来提升已有数据库系统的能力，以便满足高端用户复杂的数据分析需求。

1.4.4 信息检索

信息检索是信息按一定的方式进行加工、整理、组织并存储起来，再根据信息用户特定的需要将相关信息准确地查找出来的过程。信息检索是搜索文档或文档中信息的科学。文档可以是文本或多媒体，并且可能驻留在 Web 上。传统的信息检索与数据库系统之间的差别有两点：信息检索假定所搜索的数据是无结构的；信息检索查询主要用关键词，没有复杂的结构[不同于数据库系统中的 SQL(结构化查询语言)查询]。

信息检索的典型方法是采用概率模型。例如，文本文档可以看作词的包，即出现在文档中的词的多重集。文档的语言模型是生成文档中词的包的概率密度函数。两个文档之间的相似度可以用对应的语言模型之间的相似性度量。此外，一个文本文档集的主题可以用词汇表上的概率分布建模，称作主题模型。一个文本文档可以涉及多个主题，看作多主题混合模型。通过集成信息检索模型和数据挖掘技术，可以找出文档集中的主要主题，对集合中的每个文档，找出所涉及的主要主题。

由于万维网和电商平台、数字图书馆、数字政府、社交媒体平台等应用快速增长，大量文

本和多媒体数据日益累积并且可以联机获得。对它们的有效搜索和分析对大数据挖掘提出了许多挑战性问题。因此，文本挖掘(text mining)和多媒体挖掘与信息检索方法集成已经变得日益重要。

1.4.5　算法

与机器学习模型相辅相成的是算法及算法的实现。在大数据分析的应用中，由于海量数据会显著增加算力需求，算法设计的重要性尤为突出，对大数据分析的效率具有重要影响。如决策树分类模型就有 ID3、C4.5 和 CART(分类与回归树)等算法。

从算法的角度来看，处理大数据主要有两个思路：一是降低算法的复杂度，即减小计算量，如对社交网络或电商平台等数据量特别大的数据集，采用抽样方法，再使用随机梯度下降算法；二是分布式计算，它的基本思想是把一个大问题分解成很多小问题，然后分而治之，如 MapReduce 框架。

1.4.6　分布式计算

互联网多源异构大数据，如社交网络、搜索数据和电商平台数据，是典型的非结构化大数据，需要占用海量的存储空间和较高性能的算力，单台计算机往往很难完成数据的分析与挖掘任务，需要借助分布式计算的方式来解决。

当前流行的分布式系统 Hadoop，已经被工业界诸多企业用作大规模数据存储和处理的标准工具。Hadoop 包括两个核心工具：HDFS(Hadoop 分布式文件系统)和 MapReduce。HDFS 实现数据存储，是一个运行在普通计算机组成的集群中的分布式文件系统，适合大文件的存储和处理，能够处理 GB、TB 甚至是 PB 级别的数据，具有很强的扩展性。MapReduce 实现数据处理，数据处理流程被分解成一个个 MapReduce 作业，特别适合数据的批量处理。使用 MapReduce 完成数据分析和建模任务，需要对算法的处理逻辑和流程进行重新设计。Spark 是另一个高效的分布式计算系统，核心思想是使用内存代替磁盘作为计算过程的数据存储，大大加快数据处理速度。

课后习题

1. 简述大数据的特征。
2. 简述大数据挖掘、知识发现与商务智能之间的关系。
3. 简述大数据分析与挖掘的过程。
4. 简述大数据分析挖掘的常见模式。
5. 简述大数据分析与挖掘的技术。

应用实例

即测即练