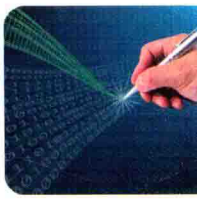
普通高等教育“十三五”应用型人才培养规划教材

数据处理与知识发现

Data Processing and Knowledge Discovery

主　编 ◎ 徐　琴　刘智珺
副主编 ◎ 王　晶
参　编 ◎ 黄向宇





机械工业出版社
CHINA MACHINE PRESS

普通高等教育“十三五”应用型人才培养规划教材

数据处理与知识发现

主　编　徐　琴　刘智珺
副主编　王　晶
参　编　黄向宇

机械工业出版社

本书系统地介绍了数据预处理、数据仓库和数据挖掘的原理、方法及应用技术，以及采用 Mahout 对相应的挖掘算法进行实际练习。本书共有 11 章，分为两大部分。第 1 ~ 7 章为理论部分。第 1 章为绪论，介绍了数据挖掘与知识发现领域中的一些基本理论、研究方法等，也简单介绍了 Hadoop 生态系统中的 Mahout；第 2 ~ 7 章按知识发现的过程，介绍数据预处理的方法和技术、数据仓库的构建与 OLAP 技术、数据挖掘原理及算法（包括关联规则挖掘、聚类分析方法、分类规则挖掘）、常见的数据挖掘工具与产品。第 8 ~ 11 章为实验部分，采用 Mahout 对数据挖掘各类算法进行实际练习。

本书应用性较强，与实践相结合，以小数据集为例详细介绍各种挖掘算法，使读者更易掌握挖掘算法的基本原理及过程；使用最广泛的大数据平台——Hadoop 生态系统中的 Mahout 对各种挖掘算法进行实际练习，实战性强，也符合目前数据处理与挖掘的发展趋势。

本书既便于教师课堂讲授，又便于自学者阅读，可作为高等院校高年级学生“数据挖掘技术”“数据仓库与数据挖掘”“数据处理与智能决策”等课程的教材。

（责任编辑邮箱：jinacmp@163.com）

图书在版编目（CIP）数据

数据处理与知识发现/徐琴，刘智珺主编．—北京：机械工业出版社，2018.8

普通高等教育“十三五”应用型人才培养规划教材

ISBN 978-7-111-60584-3

Ⅰ.①数… Ⅱ.①徐…②刘… Ⅲ.①数据处理—高等学校—教材 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字（2018）第 171144 号

机械工业出版社（北京市百万庄大街 22 号 邮政编码 100037）

策划编辑：吉 玲 责任编辑：吉 玲 范成欣 刘丽敏

责任校对：肖 琳 封面设计：张 静

责任印制：张 博

北京华创印务有限公司印刷

2018 年 9 月第 1 版第 1 次印刷

184mm × 260mm · 18 印张 · 440 千字

标准书号：ISBN 978-7-111-60584-3

定价：45.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

电话服务 网络服务

服务咨询热线：010-88379833 机 工 官 网：www.cmpbook.com

读者购书热线：010-88379649 机 工 官 博：weibo.com/cmp1952

教育服务网：www.cmpedu.com

封面无防伪标均为盗版 金 书 网：www.golden-book.com

前 言

现在的社会是一个高速发展的社会，科技发达，信息畅通，人们之间的交流越来越密切，生活也越来越方便，大数据就是这个高科技时代的产物，并且将会以更多、更复杂、更多样化的方式持续增长。大数据的复杂化和格式多样化，决定了应用服务平台中针对大数据的服务场景和类型的多样化，从而要求应用服务平台必须融合大数据技术来应对，传统的数据存储和分析技术已无法满足应用的需求。

目前行业中使用最广泛的大数据平台是基于 Apache 开源社区版本的 Hadoop 生态体系，阿里巴巴、腾讯、百度、脸书（Facebook）等国内外各大互联网公司的系统基本都采用 Hadoop 生态系统，来完成数据存储和处理。事实上，在未来 2～3 年预计有超过 50% 的大数据项目会在 Hadoop 框架下运行。

在大数据时代，大学生应具备一定的大数据处理能力。本书围绕大数据背景下的数据处理和知识发现问题，从基本概念入手，由浅入深、循序渐进地介绍了数据处理与知识发现过程中的数据预处理技术、数据仓库技术、数据挖掘的基本方法，并在最后使用最广泛的大数据平台——Hadoop 生态系统中的 Mahout 对各种挖掘算法进行实际练习，实战性强，也符合目前数据处理与挖掘的发展趋势。

目前，数据处理与知识发现及应用方法逐渐成为各高校信息类和管理类本科专业的必修内容。本书作为立足于本科教学的教材，具有如下特色：

（1）在逻辑安排上循序渐进，由浅入深，便于读者系统学习。

（2）内容丰富，信息量大，融入了大量本领域的新知识和新方法。

（3）作为教材，以小数据集为例详细介绍各种挖掘算法，使读者更易掌握挖掘算法的基本原理及过程；使用 Mahout 实践各种挖掘算法，符合大数据的发展趋势。

（4）图文并茂，形式生动，可读性强。

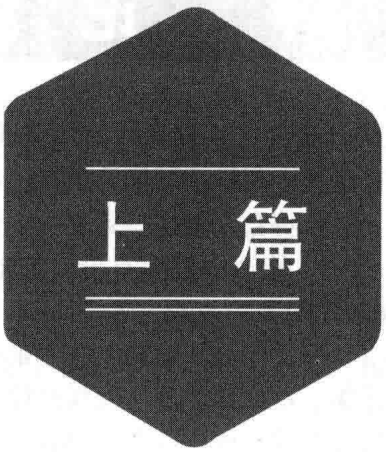
本书的编写得到了武汉民办高校合作联盟、武昌首义学院信息科学与工程学院和机械工业出版社的大力支持和帮助，在此深表谢意！

由于编者水平有限，书中难免会出现不足之处，欢迎读者批评指证。如果您有更多的宝贵意见，欢迎发邮件至邮箱 xuqin@ wsyu. edu. cn。

编 者

目 录 Contents

下篇 实验部分



理论部分

第1章 绪 论

现在的社会是一个高速发展的社会，科技发达，信息流通，人们之间的交流越来越密切，生活也越来越方便，大数据就是这个高科技时代的产物。在我们的生活中，多样化的设备和应用系统不断地产生大数据，并且将会以更多、更复杂、更多样化的方式持续增长。面对海量数据库和大量繁杂信息，如何才能从中提取有价值的知识，进一步提高信息的利用率，此为一个研究方向：基于数据库的知识发现（Knowledge Discovery in Database，KDD）以及相应的数据处理、数据挖掘（Data Mining）理论和技术的研究。

数据的处理过程可分为数据采集、数据预处理、数据存储及管理、数据分析及挖掘等环节，其中数据采集、数据存储及管理是其他课程涉及的内容，本书主要介绍数据预处理、数据分析及挖掘等内容。

本章除介绍数据仓库与数据挖掘相关的基本概念和引导性知识外，还简单介绍了 Mahout 这一基于 Hadoop 的机器学习和数据挖掘的分布式计算框架，其目的是为后续章节的学习做好基础知识的储备，并起到穿针引线的作用。

1.1 KDD 与数据挖掘

KDD 一词首次出现在 1989 年举行的第 11 届美国人工智能协会（American Association for Artifical Intelligence，AAAI）学术会议上，其后，在超大规模数据库（Very Large Database，VLDB）及其他与数据库领域相关的国际学术会议上也举行了 KDD 专题研讨会。1995 年，在加拿大蒙特利尔召开了第一届 KDD 国际学术会议（KDD'95），随后每年召开一次这样的会议。由 Kluwer Academic Publisher 出版，1997 年创刊的 Knowledge Discovery and Data Mining（知识发现和数据挖掘）是该领域中的第一本学术刊物。此后，KDD 的研究工作逐步成为热点。

知识发现和数据挖掘领域的研究工作适应市场竞争需要，它将为决策者提供重要的、潜在的信息或知识，从而产生不可估量的效益。目前，关于 KDD 的研究工作已经被众多领域所关注，如过程控制、信息管理、商业、医疗、金融等领域。

1.1.1 KDD 的定义

人们给 KDD 下过很多定义，内涵也各不相同，目前公认的定义是由美国 Microsoft Research Labs 的 Fayyad 等人提出的。基于数据库的知识发现（KDD）是指从大量数据中提取有效的、新颖的、潜在有用的、最终可被理解的模式的非平凡的过程。

数据：指一个有关事实 F 的集合，用以描述事物的基本信息。如学生学籍管理数据库

中有关学生基本情况的记录。一般来说，这些数据都是准确无误的。

模式：语言 L 中的表达式 E，E 所描述的数据是集合 F 的一个子集 F_E。F_E 表明数据集中的数据具有特性 E。作为一个模式，E 比枚举数据子集 F_E 简单。例如，“如果分数在 81 ~ 90 之间，则成绩优良”可称为一个模式。

非平凡过程：KDD 是由多个步骤构成的处理过程，包括数据预处理、模式提取、知识评估及过程优化。非平凡过程是指具有一定程度的智能性和自动性，而不仅是简单的数值统计和计算。

有效性（可信性）：从数据中发现的模式必须有一定的可信度。函数 C 将表达式映射到度量空间 M_C，c 表示模式 E 的可信度，$c = C(E,F)$，其中 $E \in L$，E 所描述的数据集合 $F_E \subseteq F$。

新颖性：提取的模式必须是新颖的。模式是否新颖可以通过以下两个途径来衡量：一是通过当前得到的数据和以前的数据或期望得到的数据之间的比较结果来判断该模式的新颖程度；二是通过对比发现的模式与已有模式的关系来判断。通常用一个函数来表示模式的新颖程度 $N(E,F)$，该函数的返回值是逻辑值或是对模式 E 的新颖程度的一个判断数值。

潜在作用：指提取出的模式将来会实际运用。通过函数 U 把 L 中的表达式映射到度量空间 M_U，u 表示模式 E 的有作用程度，$u = U(E,F)$。

可理解性：发现的模式应该能被用户理解，以帮助人们更好地了解和使用数据库中的信息，这主要体现在简洁性上。要想让一个模式更易于理解并不容易，需要对其简单程度进行度量。用 s 表示模式 E 的简单度（可理解度），它也通过函数来反映，即 $s = S(E,F)$。

上述度量函数只是从不同角度进行模式评价，往往采用权值来进行综合评判。在某些 KDD 系统中，利用函数来求得模式 E 的权值 $i = I(E,F,C,N,U,S)$；在另外一些系统中，通过对求得的模式的不同排序来表示模式的权值大小。

1.1.2　KDD 过程与数据挖掘

KDD 是一个反复迭代的人机交互处理过程。该过程需要经历多个步骤，并且很多决策需要由用户提供。从宏观上看，KDD 过程主要由以下部分组成：数据清理、数据集成、选择与变换、数据挖掘、模式评估与知识表示，如图 1-1 所示。

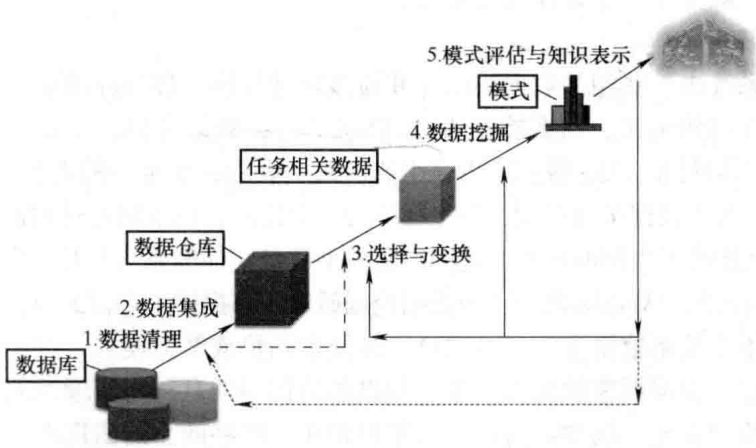


图 1-1　KDD 过程示意图

1）数据清理：消除噪声和不一致数据，如删除无效数据，用统计方法填充丢失数据等。

2）数据集成：将多种数据源的数据组合在一起，一个流行的趋势是将数据清理和数据集成作为预处理步骤执行，结果数据存放在数据仓库中。数据仓库是数据挖掘的一种对象。

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：数据变换或统一成适合挖掘的形式，如通过汇总或聚集操作。

5）数据挖掘。

① 确定 KDD 目标：首先根据用户的要求，确定 KDD 要发现的知识的类型，因为对 KDD 的不同要求会在具体的知识发现过程中采用不同的知识发现算法，如分类、关联规则、聚类等。

② 选择算法：根据确定的任务选择合适的知识发现算法，包括选取合适的模型和参数。同样的目标可以选用不用的算法来解决，这可以根据具体情况进行分析选择。选择算法的途径有以下两种：一是根据数据的特点不同，选择与之相关的算法；二是根据用户的要求，有的用户希望得到描述型的结果，有的用户希望得到预测准确度尽可能高的结果，不能一概而论。总之，要做到选择算法与整个 KDD 过程的评判标准相一致。

③ 数据挖掘：这是整个 KDD 过程中很重要的一个步骤。运用前面选择的算法，从数据中提取出用户感兴趣的数据模式，并以一定的方式表示出来是数据挖掘的目的。

6）模式评估：根据某种兴趣度度量，识别表示知识的真正有趣的模式。

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。

在上述步骤中，数据挖掘占据着非常重要的地位，它主要是利用某些特定的知识发现算法，在一定的运算效率范围内，从数据中发现出有关知识，决定了整个 KDD 过程的效果与效率。

1.2 数据挖掘的对象

数据挖掘的对象原则上可以是各种存储方式的信息。目前的信息存储方式主要包括关系数据库、数据仓库、事务数据库、高级数据库系统、文件数据和 Web 数据等，其中高级数据库系统包括面向对象数据库、关系对象数据库以及面向应用的数据库（如空间数据库、时态数据库、文本数据库、多媒体数据库等）。

1. 关系数据库

一个数据库系统由一些相关数据构成，并通过软件程序管理和存储这些数据。数据库管理系统提供数据库结构定义，数据检索语言（SQL 等），数据存储，并发、共享和分布式机制，数据访问授权等功能。关系数据库由表组成，每个表由一个唯一的表名，属性（列或域）集合组成表结构，表中数据按行存放，每一行称为一个记录，记录间通过键值加以区别。关系表中的一些属性域描述了表间的联系，这种语义模型就是实体联系（E-R）模型。关系数据库是目前最流行、最常见的数据库之一，为数据挖掘研究工作提供了丰富的数据源。

当数据挖掘用于关系数据库时，可以进一步搜索趋势或数据模式。例如，数据挖掘系统可以分析顾客数据，根据顾客的收入、年龄和以前的信用信息预测新顾客的信用风险。数据挖掘系统也可以检测偏差。例如，与以前的年份相比，哪些商品的销售出人预料。可以进一步考察这种偏差，如数据挖掘可能发现这些商品的包装的变化，或价格的大幅度提高。

2. 数据仓库

数据仓库可以把来自不同数据源的信息以同一模式保存在同一个物理地点。其构成需要经历数据清理、数据格式转换、数据集成、数据载入及阶段性更新等过程。数据仓库（Data Warehouse，DW）是一个面向主题的（Subject Oriented）、集成的（Integrated）、相对稳定的（Non-Volatile）、反映历史变化的（Time Variant）、支持管理决策（Decision Making Support）的数据集合。面向主题是指数据仓库的组织围绕一定的主题，不同于日复一日的操作和事务处理型的组织，而是通过排斥对决策无用的数据等手段提供围绕主题的简明观点。集成是指数据仓库将多种异质数据源集成为一体，如关系数据库、文件数据、在线事务记录等。数据存储包含历史信息（如过去的5～10年）。数据仓库要将分散在各个具体应用环境中的数据转换后才能使用，所以它不需要事务处理、数据恢复、并发控制等机制。

数据仓库根据多维数据库结构建模，每一维代表一个属性集，每个单元存放一个属性值，并提供多维数据视图，允许通过预计算快速地对数据进行总结。尽管数据仓库中集成了很多数据分析工具，但仍然需要像数据挖掘等更深层次、自动的数据分析工具。数据仓库的构造和使用框架如图1-2所示。

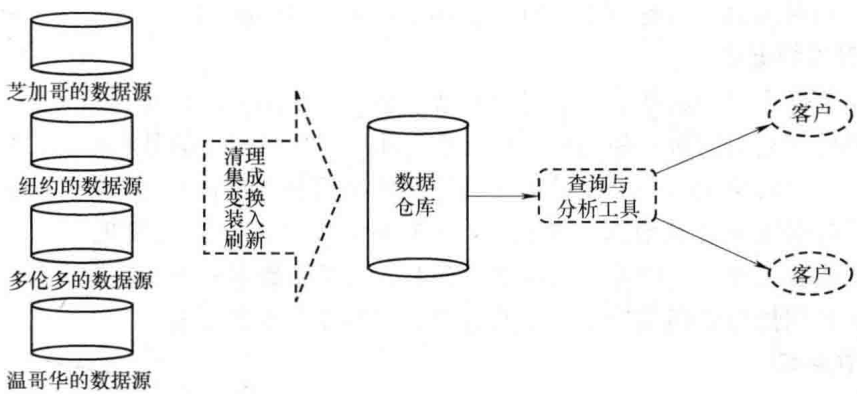


图1-2　AAA公司的数据仓库的构造和使用框架

关于数据仓库的内容主要在第3章介绍。

3. 事务数据库

一个事务数据库由文件构成，每条记录代表一个事务。通常，一个事务包含唯一的事务标识号（Trans_ID）和组成该事务的项的列表（如在超市中购买的商品）。超市的销售数据是典型的事务型数据，见表1-1。事务数据库可能有一些与之关联的附加表，如包含关于销售的其他信息：事务的日期、顾客的ID号、销售者的ID号、连锁分店的ID号等。更深层次的市场货篮（Market Basket）数据分析（如哪些商品经常同时销售等问题）只能利用数据挖掘思想来解决。

表1-1　超市销售事务数据

Trans_ID	商品ID的列表
T100	I1，I3，I8，I16
T200	I2，I8
…	…

例如，你可能问“哪些商品一起销售得很好?”，这种“购物篮数据分析”使你能够制定促销策略，将商品捆绑销售。例如，有了“打印机与计算机经常一起销售”的知识，你可以给购买指定计算机的顾客以较大的折扣（甚至免费）提供某种打印机，以期销售更多较贵的计算机（通常比打印机更贵）。传统的数据库系统不能进行购物篮数据分析。事务数据上的数据挖掘可以通过挖掘频繁项集来做这件事。频繁项集是频繁地一起销售的商品的集合。

4. 面向对象数据库

面向对象数据库是基于面向对象程序设计的范例，是面向对象程序设计技术与数据库技术结合的产物。面向对象数据库每一个实体作为一个对象，与对象相关的程序和数据封装在一个单元中，通常用一组变量描述对象，等价于实体联系模型和关系模型中的属性。对象通过消息与其他对象或数据库系统进行通信。对象机制提供一种模式获取消息并做出反应的手段。类是对象共享特征的抽象。对象是类的实例，也是基本运行实体。可以把对象类按级别分为类和子类，实现对象间属性共享。其主要特点是具有面向对象技术的封装性和继承性，提高软件的可重用性。

常见的面向对象数据库有 Object Store、Ontos、O2、Jasmin 等。

5. 关系对象数据库

关系对象数据库的构成基于关系对象模型，是对关系模型的扩充，因为大部分复杂的数据库应用需要处理复杂的对象和结构。它继承了面向对象数据库的基本概念，把每个实体看作一个对象，每个对象关联一个变量集（对应于关系模型的属性）、一个消息集（使用它可与其他对象或数据库系统其他部分通信）、一个方法集（每个方法实现一个消息的代码）。关系对象数据库在工业、应用等方面越来越普遍。与关系数据库上的数据挖掘相比，关系对象数据库上的数据挖掘更强调操作复杂的对象结构和复杂数据类型。

6. 空间数据库

空间数据库是指在关系型数据库内部对地理信息进行物理存储。常见的空间数据库数据类型包括地理信息系统、遥感图像数据、医学图像数据。空间数据可以用包括 n 维位图、像素图等光栅格式表示（如二维卫星图像数据可以用光栅格式表示，每一个像素记录一个降雨区域），也可以用向量形式表示（如道路、桥梁、建筑物等基本地理结构可以用点、线、多边形等几何图形表示为向量格式）。空间数据库具有一些共同的特点：数据量庞大、空间数据模型复杂、属性数据和空间数据联合管理、应用范围广泛。

对空间数据库可以进行何种数据挖掘呢?

例如，数据挖掘可以发现描述坐落在特定类型地点（如公园）附近的房屋特征，可能描述不同海拔的山区气候，或根据城市离主要高速公路的距离描述大城市贫困率的变化趋势。另外，可以将移动对象的趋势分组，识别移动怪异的车辆，或根据疾病随时间的地理分布，区别生物恐怖攻击与正常的流感爆发。

7. 时态数据库和时间序列数据库

时态数据库和时间序列数据库都存放与时间有关的数据。

时态数据库通常存放与时间相关的属性值，这些属性可以是具有不同语义的时间戳，如与时间相关的职务、工资等个人信息数据及个人简历信息数据等均属于时态数据库数据。

时间序列数据库存放随时间变化的值序列，如零售行业的产品销售数据、股票数据、气

象观测数据等均为时间序列数据。

对时态数据库和时间序列数据库的数据挖掘，通过研究事物发生发展的过程，可以发现数据对象的演变特征或对象变化趋势。例如，对银行数据的挖掘可能有助于根据顾客的流量安排银行出纳员；可以挖掘股票交易数据，发现可能帮助你制定投资策略的趋势，如何是实时购买某支股票的最佳时机。

8. 文本数据库

文本数据库是包含用文字描述的对象的数据库。这里的文字不是简单的关键字，可能是长句子或图形，如产品说明书、出错或调试报告、警告信息、简报等文档信息。文本数据类型包括无结构类型（大部分的文本资料和网页）、半结构类型（XML 数据）、结构类型（图书馆数据）。

通过挖掘文本数据可以发现如文本文档的简明概括的描述、关键词或内容关联，以及文本对象的聚类行为等。

9. 多媒体数据库

在多媒体数据库中主要存储图形（Graphics）、图像（Image）、音频（Audio）、视频（Video）等。多媒体数据库管理系统提供在多媒体数据库中对多媒体数据进行存储、操纵和检索的功能，特别强调多种数据间（如图像、声音等）的同步和实时处理，主要应用在基于图片内容的检索、语音邮件系统、视频点播系统。对于多媒体数据库的数据挖掘，需要将存储和检索技术相结合。目前的主要方法包括构造多媒体数据立方体、多媒体数据库的多特征提取、基于相似性的模式匹配等。

10. 万维网数据

万维网（WWW）可以被看成最大的文本数据库。万维网提供了丰富的、世界范围的联机信息服务，用户通过链接，从一个对象到另一个对象，寻找感兴趣的信息。这种系统对数据挖掘提供了大量的机会和挑战。

面向 Web 的数据挖掘比面向数据库和数据仓库的数据挖掘要复杂得多，这是由于互联网上异构数据源环境、数据结构的复杂性、动态变化的应用环境等特性所决定的。Web 数据挖掘包括 Web 结构挖掘、Web 使用挖掘、Web 内容挖掘。

例如，理解用户的访问模式不仅有助于改进系统设计（通过提供高度相关的对象间的有效访问），而且还可以导致更好的市场决策（如通过在频繁访问的文档上布置广告，或提供更好的顾客分类和行为分析等）。

11. 流数据

与传统数据库中的静态数据不同，流数据是海量甚至可能是无限，动态变化，以固定的次序流进和流出，只允许一遍或少数几遍扫描，要求快速（常是实时的）响应时间。与传统数据库相比，流数据在存储、查询、访问、实时性的要求等方面都有很大区别。

流数据的主要应用场合包括网络监控、网页点击流、股票交易、流媒体、气象或环境监控数据等。

挖掘数据流涉及流数据中的一般模式和动态变化的有效发现。例如，人们可能希望根据消息流中的异常检测计算机网络入侵，这可以通过数据流聚类、流模型动态构造或将当前的频繁模式与前一次的频繁模式进行比较来发现。

1.3 数据挖掘的任务

通常，数据挖掘任务分为以下两大类。

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目标变量（Target Variable）或因变量（Dependent Variable），而用来做预测的属性称说明变量（Explanatory Variable）或自变量（Independent Variable）。例如，用于预测离散的目标变量，如预测一个 Web 用户是否会在网上书店买书是分类任务；用于预测连续的目标变量，如预测某股票的未来价格则是回归任务；从数据集中发现与众不同的数据是离群点检测等。

典型的分类型任务如下：

- 给出一个客户的购买或消费特征，判断其是否会流失。
- 给出一个信用卡申请者的资料，判断其编造资料骗取信用卡的可能性。
- 给出一个病人的症状，判断其可能患的疾病。
- 给出大额资金交易的细节，判断是否有洗钱的嫌疑。
- 给出很多文章，判断文章的类别（如科技、体育、经济等）。

2）描述任务。通过对数据集的深度分析，寻找出概括数据相互联系的模式或规则，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。例如，把没有预定义类别的数据划分成几个合理的类别是聚类分析、任务发现数据项之间的关系是关联分析、形成数据高度浓缩的子集及描述是摘要任务等。

典型的描述型任务如下：

- 给出一组客户的行为特征，将客户分成多个行为相似的群体。
- 给出一组购买数据，分析购买某些物品和购买其他物品之间的联系。
- 给出一篇文档，自动形成该文档的摘要。

1. 关联分析

我们经常会碰到这样的问题：

① 商业销售上，如何通过交叉销售，以得到更大的收入？

② 保险方面，如何分析索赔要求，发现潜在的欺诈行为？

③ 银行方面，如何分析顾客消费行业，以便有针对性地向其推荐感兴趣的服务？

④ 哪些制造零件和设备设置与故障事件关联？

⑤ 哪些病人和药物属性与结果关联？

⑥ 哪些商品是已经购买商品 A 的人最有可能购买的？

在商业销售上，关联规则可用于交叉销售，以得到更大的收入；在保险业务方面，如果出现了不常见的索赔要求组合，则可能为欺诈，需要进行进一步的调查；在医疗方面，可找出可能的治疗组合；在银行方面，对顾客进行分析，可以推荐感兴趣的服务等。这些都属于关联规则挖掘问题。

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。关联规则挖掘的目的就在于在一个数据集中找出项之间的关系，从大量的数据中挖掘出有价值的描述数据项之间相互联系的有关知识。

关联分析挖掘的规则形式：Body⇒Head [support, confidence]。例如，buys（x, diapers）⇒buys（x, beers）[0.5%, 60%]，支持度为0.5%表示所分析的所有事务的0.5%同时购买diapers和beers。置信度60%意味着购买diapers的顾客60%也购买了beers。这个关联规则涉及单个重复的属性或谓词（即buys）。包含单个谓词的关联规则称为单维关联规则（Single-dimensional Association Rule）。去掉谓词符号，上面的规则可以简单地写成“diapers⇒beers [0.5%, 60%]”。

在典型情况下，如果关联规则满足最小支持度阈值和最小置信度阈值，则此关联规则被认为是有趣的。如果某一关联规则不能同时满足最小支持度阈值和最小置信度阈值，则它会被认为是不令人感兴趣的而被丢弃。这些阈值可以由用户或领域专家设定。

【例1-1】 关联分析。表1-2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 {Diaper}→{Beer}。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1-2 购物篮数据

TID	Items
1	Bread、Coke、Milk
2	Beer、Bread
3	Beer、Coke、Diaper、Milk
4	Beer、Bread、Diaper、Milk
5	Coke、Diaper、Milk
…	…

关联规则的挖掘将在第4章进行介绍。

2. 聚类分析

我们经常会碰到这样的问题：

- 如何通过一些特定的症状归纳某类特定的疾病？
- 谁是银行信用卡的黄金客户？
- 谁喜欢打国际长途，在什么时间，打到哪里？
- 对住宅区进行分析，确定自动提款机ATM的安放位置。
- 如何对用户WAP上网行为进行分析，通过客户分群，进行精确营销？

除此之外，促销应该针对哪一类客户，这类客户具有哪些特征？这类问题往往是在促销前首要解决的问题，对整个客户做分群，将客户分组在各自的群组里，然后对每个不同的群组采取不同的营销策略。这些都是聚类分析的例子。

不像分类和预测分析标号类的数据对象，聚类（Clustering）分析数据对象不考虑已知的类标号。一般情况下，训练数据中不提供类标号，因为开始并不知道类标号，可以使用聚类产生这种标号。聚类是按照某个特定标准（通常是某种）把一个数据集分割成不同的类，使得类内相似性尽可能地大，同时类间的区别性也尽可能地大。直观地看，最终形成的每个聚类在空间上应该是一个相对稠密的区域。可见，最大化类内部的相似性、最小化类之间的相似性是聚类的原则。

聚类方法主要包括划分聚类、层次聚类、基于密度的聚类、基于网格的聚类、基于模型的聚类等。

作为一种数据挖掘功能，聚类分析也可以作为一种独立的工具，用来洞察数据的分布，观察每个簇的特征，将进一步分析集中在特定的簇集合上。另外，聚类分析可以作为其他算法（如特征化、属性子集选择和分类）的预处理步骤，之后这些算法将在检测到的簇和选择的属性或特征上进行操作。例如，“哪一种类的促销对客户响应最好?”，对于这一类问题，首先对整个客户做聚集，将客户分组在各自的聚集里，然后对每个不同的聚集回答问题，可能效果更好。

【例 1-2】 聚类分析。设有记录了 4 个顾客 3 个信息的数据库，见表 1-3。

表 1-3 计算机商店顾客信息

顾客 ID	学　生	年龄段/岁	收　入	类　别
X_1	否	31～40	一般	?
X_2	是	≤30	一般	?
X_3	是	31～40	较高	?
X_4	否	≥41	一般	?

将记录进行聚类分析。由于没有指定具体的相似度标准，因此根据表 1-3 的属性，可以考虑选择几个不同的标准来进行聚类分析，并对结果进行比较。

① 以是否为“学生”为相似度标准，则 4 条记录可聚成以下两个簇：

$$A_{学生}=\{x_1,x_4\},\ B_{非学生}=\{x_2,x_3\};$$

② 以顾客的年龄段作为相似度标准，则 4 条记录可聚成以下 3 个簇：

$$A_{\leqslant 30}=\{x_2\},B_{31\sim 40}=\{x_1,x_3\},\ C_{\geqslant 41}=\{x_4\};$$

③ 以收入水平作为相似度标准，则 4 条记录可聚成以下两个簇：

$$A_{一般}=\{x_1,x_2,x_4\},B_{较好}=\{x_3\};$$

通过此例可以发现，对顾客记录的聚类分析是对顾客集合的一个恰当的划分。对一个给定顾客数据库，如果相似性度量标准不同，则划分结果也不同，即聚类算法对相似性度量标准是敏感的。这也告诉我们，可选择不同的度量标准对数据库记录进行聚类分析，以期得到更加符合实际工作需要的聚类结果。

聚类分析将在第 5 章进行介绍。

3. 分类分析

分类分析（Classification Analysis）通过分析已知类别标记的样本集合（示例数据库）中的数据对象（记录），为每个类别做出准确的描述，或建立分类模型，或提取出分类规则（Classification Rules），然后用这个分类模型或规则对样本集合以外的记录进行分类。

分类预测导出的模型的表示形式有分类（IF-THEN）规则、决策树、数学公式或神经网络，如图 1-3 所示。在图 1-3 中，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，每个分支代表测试的一个输出，而树叶代表类或类分布，决策树容易转换成为分类规则；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。

另外，还有构造分类模型的其他方法，如朴素贝叶斯分类、支持向量机和 k 最近邻分类。