高等教育规划教材

数据挖掘原理、算法与应用

梁亚声 徐欣 等编著

免费提供电子教案

下载网址 http://www.cmpedu.com

机械工业出版社
CHINA MACHINE PRESS

高等教育规划教材

数据挖掘原理、算法与应用

梁亚声　徐　欣　成小菊　梁佳领　朱　霞　编著

机械工业出版社

本书系统介绍了数据挖掘原理、算法和应用的相关知识。主要内容包括数据挖掘的过程、数据存储技术、数据预处理技术和算法、异常数据检测技术和算法、数据分类算法和应用、数据聚类分析的算法及其应用、数据关联分析算法及其应用、模型的评估技术和算法、复杂数据类型的数据挖掘技术。本书涵盖了数据挖掘过程的各方面技术和算法，在内容安排上将理论知识和工程技术应用有机地结合起来，并介绍了许多数据挖掘的典型应用方法。

本书可作为高等院校计算机科学与技术、信息管理、数据分析等专业的教科书，也可作为企业管理、信息分析人员的技术参考书。

本书配有电子课件，需要的教师可登录 www. cmpedu. com 免费注册，审核通过后下载，或联系编辑索取（QQ：2399929378，电话：010 - 88379753）。

图书在版编目(CIP)数据

数据挖掘原理、算法与应用/ 梁亚声等编著. —北京：机械工业出版社，2014. 11
高等教育规划教材
ISBN 978-7-111-49632-8

Ⅰ. ①数… Ⅱ. ①梁… Ⅲ. ①数据采集 - 高等学校 - 教材
Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2015）第 049163 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：郝建伟
责任编辑：郝建伟 吴晋瑜 责任校对：张艳霞
责任印制：李 洋
北京宝昌彩色印刷有限公司印刷
2015 年 4 月第 1 版 · 第 1 次印刷
184mm × 260mm · 20. 75 印张 · 513 千字
0001 - 3000 册
标准书号：ISBN 978-7-111-49632-8
定价：49. 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
服务咨询热线：(010)88379833
读者购书热线：(010)88379649

网络服务
机 工 官 网：www. cmpbook. com
机 工 官 博：weibo. com/cmp1952
教育服务网：www. cmpedu. com
金 书 网：www. golden - book. com

出 版 说 明

当前，我国正处在加快转变经济发展方式、推动产业转型升级的关键时期。为经济转型升级提供高层次人才，是高等院校最重要的历史使命和战略任务之一。高等教育要培养基础性、学术型人才，但更重要的是加大力度培养多规格、多样化的应用型、复合型人才。

为顺应高等教育迅猛发展的趋势，配合高等院校的教学改革，满足高质量高校教材的迫切需求，机械工业出版社邀请了全国多所高等院校的专家、一线教师及教务部门，通过充分的调研和讨论，针对相关课程的特点，总结教学中的实践经验，组织出版了这套“高等教育规划教材”。

本套教材具有以下特点：

1）符合高等院校各专业人才的培养目标及课程体系的设置，注重培养学生的应用能力，加大案例篇幅或实训内容，强调知识、能力与素质的综合训练。

2）针对多数学生的学习特点，采用通俗易懂的方法讲解知识，逻辑性强、层次分明、叙述准确而精练、图文并茂，使学生可以快速掌握，学以致用。

3）凝结一线骨干教师的课程改革和教学研究成果，融合先进的教学理念，在教学内容和方法上做出创新。

4）为了体现建设“立体化”精品教材的宗旨，本套教材为主干课程配备了电子教案、学习与上机指导、习题解答、源代码或源程序、教学大纲、课程设计和毕业设计指导等资源。

5）注重教材的实用性、通用性，适合各类高等院校、高等职业学校及相关院校的教学，也可作为各类培训班教材和自学用书。

欢迎教育界的专家和老师提出宝贵的意见和建议。衷心感谢广大教育工作者和读者的支持与帮助！

机械工业出版社

前　言

随着信息技术的普及和应用，各个领域产生了大量的数据，这些数据被获取、存储下来，其中蕴含着丰富的信息。人们持续不断地探索处理这些数据的方法，以期最大程度地从中挖掘有用的信息，面对如潮水般不断增加的数据，人们不再满足于数据的查询和统计分析，而是期望从数据中提取信息或者知识为决策服务。数据挖掘技术突破了数据分析技术的种种局限，它结合统计学、数据库、机器学习等技术解决从数据中发现新的信息，辅助决策这一难题，是正在飞速发展的前沿学科。一些大型企业对数据挖掘产品和工具的使用都超过 20 年，并已产生了期望的效应。此外，数据挖掘产品和工具在金融、商业、电信、医学等多个领域也得到广泛推广应用。

数据挖掘并不是要取代其他数据分析技术，而是将它们作为其工作的基础。尽管有些技术（如关联分析）是数据挖掘独有的，但是，另一些技术（如聚类、分类和异常检测）则建立在其他学科长期研究的基础之上。数据挖掘利用已有技术加速其发展，并一直与其他学科的技术紧密结合。成功地进行数据挖掘是综合使用多种技术，以及理解数据的专业人员和数据分析人员合作的成果。

本书结合数据挖掘技术的最新发展，系统地介绍了数据挖掘的基础理论、技术原理、算法和应用，以使读者对数据挖掘有一个系统、全面的了解。本书共 9 章，第 1 章主要介绍数据挖掘的基本概念和数据挖掘的过程。第 2 章主要介绍关系数据集和数据仓库等数据存储方式的基本概念、数据组织及其涉及的关键技术，以及分布式文件系统、NoSQL 等大数据存储方式的概念、结构、原理和数据组织方法等。第 3 章主要介绍了数据预处理的概念和必要性，以及数据清理、数据集成、数据转换、数据归约、数据离散化和特征选择等数据预处理技术。第 4 章主要介绍了相似度度量的基础知识和 5 种异常检测方法，并深入分析欧式距离等 6 种传统的度量方法和大数据度量方法。第 5 章主要介绍了数据分类和预测的基本概念，决策树分类、贝叶斯分类、神经网络等分类方法，以及预测算法与应用。第 6 章主要介绍了数据聚类分析的基本概念，以及基于划分、基于层次、基于密度、基于网格和基于模型的聚类算法，还介绍了聚类分析的评估方法及其应用。第 7 章主要介绍了关联分析的基本概念，分析了关联规则挖掘的两个子任务：频繁项集产生和规则产生，介绍了频繁项集的紧凑表示及产生频繁项集的其他方法、FP – growth 算法、关联评估及其应用等内容。第 8 章主要针对数据挖掘模型的评价和度量介绍了评分函数（包括常用的预测性评分函数和描述性评分函数）；介绍了针对数据挖掘模型

的成本评价曲线；从评价模型复杂度角度介绍了最短描述长度原则等评价方法；针对模型有效性验证介绍了交叉验证和 Bootstrap 验证方法；从数据挖掘模型效率和准确率提升角度，介绍了云计算和集成学习方法。第 9 章主要介绍了针对文本、图像、语音识别、视频、网络拓扑、网络舆情、推荐系统、空间数据和数据流等复杂数据的数据挖掘技术，分析了各类复杂数据的特点，介绍了相关数据挖掘的关键技术。本书涉及的内容较为广泛，在教学时，可根据实际情况选择。

本书由梁亚声编写第 1、5 章，徐欣编写第 8、9 章，成小菊编写第 6、7 章，梁佳领编写第 2、3 章，朱霞编写第 4 章。何成宇为第 4 章的编写提供了部分资料。徐欣对全书进行了统稿。

本书编著得到了国家自然科学基金（61402426）资助。

由于作者水平有限，书中难免存在不妥之处，敬请读者批评指正。

编　者

目　录

第1章 概 述

随着信息化的普及，各领域都积累和收集了大量数据，这些数据涵盖了商业、科技、政治等各类重要信息。但是，面对庞大的数据资源，人们所使用的只是其中一小部分。而借助数据挖掘技术，人们可以从浩瀚的数据海洋中，挖掘出有价值的信息和知识，以作决策支持之用。

1.1 从数据中获取知识

知识是人类对客观世界的观察和了解，是人类在实践中认识客观世界的成果。知识推动人类的进步和发展。人类所作出的正确判断和决策以及采取正确的行动都基于智慧和知识。在信息化的现代社会中，知识在各个方面都占据着中心地位，并起着决定性的作用。知识是事物的概念或规律，源于外部世界，所以知识是客观的。但是知识本身并不是客观现实，而是事物的特征与联系在人脑中的反映。

数据是反映客观事物的数字、词语、声音和图像等，是可以进行计算加工的“原料”。数据是对客观事物的数量、属性、位置及其相互关系的抽象表示，适于存储、传递和处理。随着信息技术的发展，每天数以亿计的海量数据被获取、存储和处理。这些海量数据蕴含着大量的信息、潜在的规律或规则。人们可以通过海量数据了解客户的需求、预测市场动向等。然而，数据仅仅是人们运用各种工具和手段观察外部世界所得到的原始材料，从数据到知识再到智慧，需要经过分析、加工、处理和精炼等一系列过程。

“啤酒与尿布”是沃尔玛利用数据获取知识的成功案例。1983 年，沃尔玛借助信息技术发明了条形码、无线扫描枪、计算机跟踪存货等新技术，使各部门、各业务流程运行得迅速、准确。同时，数据库系统中积累了包括大量顾客消费行为记录在内的海量经营数据。沃尔玛在对海量数据进行分析时意外发现：“跟尿布一起购买最多的商品是啤酒”。经过深入研究，人们发现这些数据揭示了“尿布与啤酒”这一现象背后所隐藏的美国人的一种行为模式，即年龄在 25 ~ 35 岁的年轻父亲下班后经常要到超市去给婴儿买尿布，其中 30% ~ 40% 的人会顺手买几瓶啤酒。沃尔玛立即采取了行动，将卖场内原本相隔很远的妇婴用品区与酒类饮料区的空间距离缩短，使顾客更加方便，然后对新生育家庭的消费能力进行了调查，对这两个产品的价格也做了调整，结果使尿布与啤酒的销售量大增。

随着计算机技术、数据库技术、传感器技术和自动化技术的飞速发展，数据的获取、存储变得越来越容易。这些数据和由此产生的信息如实地记录着事物的本质状况。但是海量数据的激增迫使人们不断寻找新的工具，以满足其对规律进行探索，进而为决策提供有效信息的需求。

1.2 数据挖掘的基本概念

数据挖掘是一种信息处理技术，是从大量数据中自动分析并提取知识的技术。数据挖掘是一个处理过程，是从大量数据中挖掘出隐含的、先前未知的、对决策有价值的知识的过程。数据挖掘的目的是从所获取的数据中发现新的、规律性的信息和知识，以辅助科学决策，利用各种分析工具对海量数据进行深入归纳、分析，从而获得对所研究对象更深层次的认识，发现隐藏在数据中的数据之间规律性的关系、发现可以预测趋势的数学模型，并用这些知识和规则建立用于决策支持的模型，用来分析风险、进行预测。

数据挖掘是通过仔细分析大量数据来揭示有意义的新的关系、模式和趋势的过程。它使用的技术包括模式认知技术、统计技术和数学技术。数据挖掘所获取的知识是以模型或数据概化的形式给出的。数据挖掘技术有许多种类，其方法都采用基于归纳的学习。基于归纳的学习是通过观察所学概念的特定实例形成一般概念的过程，例如，信用卡公司记录信用卡使用者购买习惯的常用模型，当其交易不符合习惯模型时，将怀疑信用卡已被盗用。

数据挖掘可以从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中、人们事先不知道的、但又可能有用的信息和知识。一般情况下，数据挖掘的算法大多建立在统计学大数定律基础上。若数据量太小，则常常无法反映出真实世界中的普遍特性，利用挖掘算法所得出的结论也不可靠。但并非小数据量就不可以进行挖掘，近年来，研究者也提出了一些对小样本进行挖掘的方法，如支撑向量机方法就是基于小样本学习理论的非常实用的方法。数据量虽小，但数据总是事物特性一定程度的反映，只要建立的模型和算法得当，当然也可以从这些数据中获取一定的信息。

从理论上说，数据量越大越好。但随着数据量的增大，算法执行效率会越来越低，甚至无法计算。

在现实世界中，所获取数据往往具有不完全、有噪声、模糊、随机性等特点，例如，进行问卷调查时，不少人不填婚姻状况和年龄，这些不完全或缺失的数据会给数据挖掘增加一定的难度，进行数据挖掘时，我们可以删除这些样本或记录，也可以采用一定的方法将这些缺失数据补上，或者使用可以自动处理缺失数据的算法。

在实际工作中，还会遇到异常情况的干扰，使获得的数据偏离了真实值，这样的数据就是噪声数据。不光外界的干扰，测量仪器的故障、人工输入或抄写所致的失误等都可能形成噪声数据，实际问题中噪声数据往往难以避免。这些问题都需要在数据挖掘过程中予以解决。

1.3 数据挖掘的发展历程

20 世纪 60 年代，人们借助计算机以文件方式对数据进行管理。在 70 年代，关系数据库的发明和使用，使人们能够收集、存储、处理大量的数据。随着计算机网络的应用，人们逐渐采用联机事务处理（OnLine Transaction Processing，OLTP）对信息数据进行及时、高效的存储和处理。OLTP 重点在于业务操作，对当前数据进行及时处理。

提供决策支持建立在对大量历史数据进行分析处理的基础上。为了避免长时间占用系统资源，影响日常数据实时处理，我们需要把相关数据从事务处理系统中提取出来，并按照决策支持的需要进行重新组织，建立相应的分析环境。W. H. Inmon 于 1993 年出版了《Building the

Data Warehouse》，给出了数据仓库定义：数据仓库是一个面向主题的、集成的、稳定的、不同时间的数据集合，用于支持管理层的决策过程。

与数据仓库同时产生的还有对数据进行汇集、合并和聚集以及从不同角度观察信息的分析技术，即联机分析处理（OnLine Analytical Processing，OLAP）。人们通过 OLAP 技术可以对从数据库或数据仓库得到的经验、规则进行验证，也可以对数据挖掘结果的有效性进行检验和完善。

然而，数据库和数据仓库越建越大，通过直观的感觉、简单的统计分析和 OLAP 技术并不能完全发现隐藏在数据中有价值的信息和知识。“沉浸在数据的海洋中，却渴望着知识的淡水”这句话生动地描绘了人们面对海量数据的迷惘和无奈。沃尔玛“啤酒与尿布”的故事使人们看到了数据挖掘的作用。人们试图利用数据挖掘来解决所面临的问题，如客户分群、客户流失原因及预测、关联消费、指导生产和管理等。人们希望将商业管理、生产控制、市场分析、工程设计、科学探索等海量数据资源转换为信息和知识。20 世纪 90 年代中期以后，基于数理统计、人工智能、机器学习、神经网络等多种技术的出现和发展使关于数据挖掘软件的开发和应用成为热点。

在 1989 年 8 月第 11 届国际人工智能联合会议上，数据挖掘的概念被正式提出，即数据库中的知识发现（Knowledge Discovery in Database，KDD）。20 世纪 90 年代开始，学术界习惯沿用 KDD 这个术语，而在商用领域，因为“数据库中的知识发现”一词过于冗长，就普遍采用了更加简单的术语——“数据挖掘”。数据挖掘系统的发展见表 1-1。

表 1-1　数据挖掘系统的发展

	特　征	数据挖掘算法	集　成	分布计算模型	数据模型
第一代	数据挖掘作为一个独立的应用	支持一个或多个算法	单独的系统	单机	向量数据
第二代	与数据库以及数据仓库集成	多个算法：能够挖掘无法一次放进内存的数据	数据管理系统，包括数据库和数据仓库	同质/局部区域的计算机群集	有些系统支持对象、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	数据管理和预言模型系统	Internet/Extranet 网络计算	支持半结构化数据和 Web 数据
第四代	与移动数据/各种计算数据联合	多个算法	数据管理、预言模型、移动系统	移动和各种计算设备	普遍存在的计算模型

第一代数据挖掘系统支持一个或少数几个数据挖掘算法，这些算法被用来挖掘向量数据（Vector - valued Data），这些数据模型在挖掘时，直接将需要挖掘的数据一次性调入内存。这些系统的成功依赖于数据的质量。许多这样的系统已经商业化。

第二代数据挖掘系统支持数据库和数据仓库，具有高性能的接口，具有高的可扩展性，例如，第二代系统能够挖掘大数据集、更复杂的数据集以及高维数据。这一代系统通过支持数据挖掘模式（Data Mining Schema）和数据挖掘查询语言（DMQL）增加系统的灵活性。第二代数据挖掘系统提供数据仓库和数据挖掘系统之间的有效接口。但是，数据仓库的设计目的是为了便于 OLAP 操作，而不是数据挖掘应用。这意味着第二代数据挖掘系统必须使用专门的数据管理系统，以此弥补数据库及数据仓库管理系统的缺陷。

第三代数据挖掘系统的特征是能够挖掘 Internet/Extranet 的分布式和高度异质的数据。这一代数据挖掘系统的关键技术之一是提供对建立在异质系统上的多个预言模型以及管理预言模型的元数据。第三代数据挖掘系统提供数据挖掘系统和预言模型系统之间的有效接口。在实施

策略方面，如果使用多个预言模型，或者预言模型需要经常修改，那么应该选择第三代数据挖掘系统。第三代数据挖掘系统也能与数据库或数据仓库集成。

第四代数据挖掘系统能够挖掘嵌入式系统、移动系统和计算设备产生的各种类型的数据。随着移动计算的重要性日渐增加，第四代数据挖掘系统起到了关键的作用。

第一代数据挖掘系统还在发展，第二、三代以及第四代数据挖掘系统将与数据仓库合并，提供集成的系统用于各种应用。并且第二、三、四代数据挖掘技术将与各种应用集成，成为一种嵌入式技术。

1.4 数据挖掘的功能和数据挖掘系统的分类

数据挖掘技术的基本功能主要体现在分类与回归、聚类分析、关联规则、时序模式、异常检测等五个方面。尽管数据挖掘技术能够增强信息检索系统的能力，但是，利用数据的明显特征来创建索引结构，查找数据库中的个别记录以及通过因特网的搜索引擎查找特定的 Web 页面，这些均不属于数据挖掘的范畴。

1.4.1 分类与回归

分类与回归主要用于解决下列的问题：

① 将信用卡申请人分为低、中、高风险群。

② 预测哪些顾客在未来半年内会取消该公司的服务，哪些电话用户会申请增值服务。

③ 预测具有某些特征的顾客是否会购买一台新的计算机。

④ 预测病人应当接受 3 种具体治疗方案中的哪一种。

⑤ 预测一位顾客在一次销售期间将花多少钱。

⑥ 预测银行可以安全地贷给贷款人的贷款量。

⑦ 预测哪些使用 2G 通信网络的手机用户有可能转换到 3G 通信网络。

⑧ 预测房地产开发中存在的风险。

分类（Classification）是构造一个分类函数（分类模型），把具有某些特征的数据项映射到某个给定的类别上。因为在分析测试数据之前，类别就已经确定了，所以分类通常被称为“有监督的学习”。分类算法要求基于数据属性值来定义类别，通常通过已知所属类别的数据的特征来描述类别。分类过程由两步构成：模型创建和模型使用。模型创建是指通过对训练数据集的学习来建立分类模型；模型使用是指使用分类模型对测试数据和新的数据进行分类。其中的训练数据集是带有类标号的，也就是说，在分类之前，要划分的类别是已经确定的。通常分类模型以分类规则、决策树或数学表达式的形式给出。

1.4.2 聚类分析

聚类分析主要用于解决下列问题：

① 通过一些特定的症状归纳某类特定的疾病。

② 预测谁是银行信用卡的黄金客户。

③ 预测谁喜欢打国际长途，在什么时间，打到什么地方。

④ 对住宅区进行聚类，确定自动提款机 ATM 的安放位置。

⑤ 对用户 WAP 上网行为进行分析，通过对客户分群进行精确营销。

聚类（Clustering）与分类不同，聚类分析是在没有给定划分类的情况下，根据信息相似度进行信息聚类的一种方法，故聚类又称为“无监督的学习”。聚类就是将数据划分或分割成相交或者不相交的群组的过程。通过确定数据之间在预先指定的属性上的相似性就可以完成聚类任务。聚类的输入是一组未被标记的数据，根据数据自身的距离或相似度进行划分。划分的原则是保持最大的组内相似性和最小的组间相似性，即使得不同聚类中的数据尽可能地不同，而同一聚类中的数据尽可能地相似，比如根据股票价格的波动情况，股票可以被分成不同的类，总共可以分成几类，各类包含哪些股票，每一类的特征是什么，这对投资者尤其对投资基金者来说，可能就是很重要的信息。当然，聚类除了将样本分类外，还可以完成孤立点挖掘，如其在网络入侵检测或金融风险欺诈探测中的应用。

1.4.3 关联规则

关联规则主要用于解决下列问题：

① 商业销售方面：如何通过交叉销售得到更大的收入。

② 保险方面：如何分析索赔要求，以发现潜在的欺诈行为。

③ 银行方面：如何分析顾客消费行为，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。

这些都属于关联规则挖掘问题，关联规则挖掘的目的就是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。随着收集和存储在数据库中的数据规模越来越大，人们可以从数据中挖掘出相应的关联知识。

关联规则（Association）揭示了数据之间的相互关系，而这种关系没有在数据中直接表示出来。关联分析的任务就是发现事物间的关联规则（或称相关程度）。

关联规则的一般形式是：若 A 发生，则 B 有百分之 C 的可能发生。C 被称为关联规则的置信度（Confidence）。

关联分析用以寻找数据库中大量数据的相关联系，其常用的两种技术为关联规则和序列模式。利用关联规则可以发现一个事物与其他事物间的相互关联性或相互依赖性，如分析客户在超市既买牙刷又买牙膏的可能性；序列模式则将重点放在分析数据之间的前后因果关系，如买了计算机的顾客会在 3 个月内买杀毒软件。

1.4.4 时序模式

时序模式主要用于解决下列问题：

① 预测下个月的商品销量、销售额或库存量。

② 预测明天广州市的最高用电负荷。

时序模式用于描述基于时间或其他序列的经常发生的规律或趋势，并对其建模。与回归一样，它也是用已知的数据预测未来的值，但这些数据的区别是变量所处时间的不同。时序模式重点考虑数据之间在时间维度上的关联性。时序模式包含时间序列分析和序列发现。

时间序列分析（Time Series）用已有的数据序列预测未来。在时间序列分析中，数据的属性值是随着时间不断变化的。回归不强调数据间的先后顺序，而时间序列要考虑时间特性，尤其要考虑时间周期的层次，如天、周、月、年等，有时还要考虑日历的影响，如节假日等。

序列发现用于确定数据之间与时间相关的序列模式。这些模式与在数据（或者事件）中发现的相关的关联规则很相似，只是这些序列是与时间相关的。

1.4.5 异常检测

异常是对差异和极端特例的表述，如分类中的反常实例、聚类外的离群值、不满足规则的特例等。大部分数据挖掘方法都将这种差异信息视为噪声而丢弃，然而在一些应用中，罕见的数据可能比正常的数据更有用。

异常检测（Outlier Detection），也被称为离群点检测，是用来发现与正常情况不同的异常和变化，并进一步分析这种变化是有意的诈骗行为，还是正常的变化。若是异常行为，则需提示预防措施，尽早防范。

1.4.6 数据挖掘系统的分类

根据不同标准，数据挖掘系统可以按不同的方式进行分类。

1. 根据数据源类型分类

针对不同的数据源，数据挖掘需要相应的挖掘技术，例如，根据数据模型，可以有关系的、事务的、对象-关系的或数据仓库的挖掘系统。如果根据所处理数据的特定类型，可以有空间的、时间序列的、文本的、流数据的、多媒体的数据挖掘系统以及 WWW 网页挖掘系统。

2. 根据数据挖掘的功能分类

数据挖掘系统可以根据挖掘的功能分类，如特征提取、区分、关联和相关分析、分类、预测、聚类、异常检测和演变分析。一个综合的数据挖掘系统通常提供多种数据挖掘功能。此外，数据挖掘系统还可以根据所挖掘知识的粒度或抽象层进行区分，包括广义知识（高抽象层）、原始层知识（原始数据层）或多层知识（考虑若干抽象层）。高级数据挖掘系统应当支持多抽象层的知识发现。数据挖掘系统还可以分为挖掘数据的规则性（通常出现的模式）系统与挖掘数据的奇异性（如异常或离群点）系统。一般来讲，概念描述、关联和相关分析、分类、预测和聚类等挖掘任务属于数据的规则性，而离群点被作为噪声排除。当然，这些方法也能帮助检测离群点。

3. 根据所用的技术分类

数据挖掘系统还可以根据用户交互程度（如自动系统、交互探查系统、查询驱动系统）以及所用的数据分析方法（如面向数据库或面向数据仓库的技术、机器学习、统计学、可视化、模式识别、神经网络等）进行分类。复杂的数据挖掘系统通常采用多种数据挖掘技术，或采用有效的、集成的技术，结合一些方法的优点。

4. 根据其应用分类

数据挖掘系统也可以根据其应用进行分类，例如，有的数据挖掘系统特别适合金融、电信、股票市场、E-mail 等。针对不同的应用，系统通常需要集成对该应用特别有效的方法。一般情况下，泛化的、全能的数据挖掘系统可能并不适合特定领域的挖掘任务。

1.5 数据挖掘的过程

数据挖掘的过程会随应用领域的不同而有所变化。每一种数据挖掘技术也有其各自的特性和使用步骤，针对不同问题和需求所制定的数据挖掘过程也会存在差异。此外，数据的完整程

度、专业人员的支持程度等都会对数据挖掘的过程有所影响。这些因素造成了数据挖掘在不同领域中的运用、规划以及流程的差异性，即使在同一领域，数据挖掘也会因为分析技术和专业知识涉入程度的不同而不同。

1.5.1 数据挖掘的一般流程

数据挖掘的一般流程可以分为明确问题、数据收集和预处理、数据挖掘以及结果解释和评估。

1. 明确问题

数据挖掘的首要工作是研究发现何种知识，即明确问题。在此过程中，数据挖掘人员必须和领域专家紧密协作，一方面明确实际工作对数据挖掘的要求；另一方面通过对各种学习算法的对比进而确定可用的学习算法（后续的学习算法选择和数据集准备都是在此基础上进行的）。

例如，数据分析员面对客户的流失问题，需要利用数据分析找出原因，并且找出解决问题的办法。

2. 数据收集和预处理

数据收集和预处理阶段一般要完成3项工作：数据选取、数据预处理和数据变换。数据选取就是确定操作对象，即目标数据，一般是从原始数据库中抽取的组数据。数据预处理一般包括消除噪声、推导计算缺失值数据、消除重复记录、完成数据类型转换（如把连续值数据转换为离散型的数据，以便用于符号归纳，或是把离散型的转换为连续值型的，以便用于神经网络）等内容。当数据挖掘的对象是数据仓库时，一般来说，数据预处理已经在生成数据仓库时完成了。数据变换的主要目的是消减数据维数，即从初始特征中找出真正有用的特征，以减少数据挖掘时要考虑的特征或变量个数。

在进行数据挖掘技术的分析之前，我们还有许多准备工作要完成，通常有80%的时间和精力花费在数据预处理阶段。数据挖掘通常有3种访问数据的途径：

- 从数据仓库中访问数据。
- 从关系数据库中访问数据。
- 从简单文件或电子表格中访问数据。

就数据仓库而言，一般情况下，数据组合的数据来源于一个或多个操作型数据库。操作型数据库是基于事务的，往往采用关系数据库模型进行设计。使用关系模型的操作型数据库包含若干个规范化的表。这些表进行了规范化以减少冗余，并加快了对记录的访问，例如，一个特定客户的数据可能在几个关系表中出现，每个表反映从不同角度看待客户的数据。

如果没有数据仓库，我们可以使用数据库查询语言（如SQL）书写查询语句，构建适合数据挖掘的表。无论挖掘的数据是从数据仓库中提取，还是通过查询语言提取，都可能需要一个实用程序将提取出的数据转换为所选数据挖掘工具要求的格式。如果还没有设计存储数据的数据库结构，并且收集的数据量是较小的，那么我们可以将数据存储在一个平面文件或电子表格中。

数据挖掘需要访问的数据可以是包含在几个数据库文件的大量记录，也可以是包含在一个文件中的几百条记录。一种普遍存在的误解是，为了建立数据挖掘算法必须具备上万条实例。实际上，即便只有几百条或几千条相关记录，大多数数据挖掘工具也能工作得很好。例如，数据分析员为了找出流失客户的原因，就要收集公司内部和外部的大量数据，包括有关持有保险