高等学校计算机科学与技术教材

- 原理与技术的完美结合
- 教学与科研的最新成果
- 语言精炼，实例丰富
- 可操作性强，实用性突出

数据挖掘
理论与应用

□ 胡可云　田凤占　黄厚宽　编著

清华大学出版社　●　北京交通大学出版社

高等学校计算机科学与技术教材

数据挖掘理论与应用

胡可云　田凤占　黄厚宽　编著

清华大学出版社
北京交通大学出版社
·北京·

内 容 简 介

本书从数据挖掘理论与数据挖掘应用过程两方面介绍了数据挖掘的最新成果。在理论部分，本书介绍了数据挖掘技术所涉及的基本概念、主流技术和最新成果；在应用部分，本书结合具体的实例系统论述了商业理解、数据预处理、建模、模型部署等整个数据挖掘流程。

本书既可以作为大学本科生和研究生的补充教材，也可以作为企业实施数据挖掘和商务智能的实战指导；既可以作为初次接触数据挖掘技术的入门读物，也可以作为高级研究人员的参考书。

图书在版编目(CIP)数据

数据挖掘理论与应用/胡可云，田凤占，黄厚宽编著. —北京：清华大学出版社；北京交通大学出版社，2008.4
（高等学校计算机科学与技术教材）
ISBN 978－7－81123－252－3

Ⅰ. 数…　Ⅱ. ① 胡…　② 田…　③ 黄…　Ⅲ. 数据采集－高等学校－教材　Ⅳ. TP274

中国版本图书馆 CIP 数据核字（2008）第 041221 号

策划编辑：谭文芳
责任编辑：郭东青　　特邀编辑：高振宇
出版发行：清 华 大 学 出 版 社　　邮编：100084　　电话：010－62776969　　http://www.tup.com.cn
　　北京交通大学出版社　　邮编：100044　　电话：010－51686414　　http://press.bjtu.edu.cn
印 刷 者：北京东光印刷厂
经　　销：全国新华书店
开　　本：185×260　　印张：16.25　　字数：399 千字
版　　次：2008 年 4 月第 1 版　　2008 年 4 月第 1 次印刷
书　　号：ISBN 978－7－81123－252－3/TP·408
印　　数：1～4 000 册　　定价：24.00 元

前　言

在企业应用中，数据挖掘这个词被越来越多地提起。企业在建立越来越多的数据库，日常运营网络化，收集越来越多的数据之后，对数据进行深度分析的需求渐渐多了起来。数据挖掘能够帮助企业从以生产为导向的企业转型为以市场为导向的企业，在日益激烈的市场竞争中立于不败之地。

数据挖掘是一门面向应用的新兴学科分支，在过去的几年中，在许多领域的应用取得了成功。特别是在计算机应用起步较早、积累了大量数据的行业，如电信、银行、零售、科学研究等。当然，数据挖掘的应用范围远不止于此。数据挖掘的应用范围极其广泛，限制数据挖掘应用范围的只是可用的数据和人们的想像力。

虽然现在很多人都认识到了数据挖掘的作用，然而在实际的数据挖掘实践中，人们往往会陷入过分强调理论或轻视业务的误区。事实上，正确的数据挖掘过程和数据挖掘理论的运用对实践的数据挖掘项目的成功同等重要。本书编写的目的，是既完整地介绍数据挖掘常用的算法，又对数据挖掘的应用过程进行详细的介绍，使读者能够掌握整个数据挖掘过程的各个方面，从而更好地进行数据挖掘实践。

本书的特色是理论与应用相结合。在充分介绍数据挖掘的主流技术和最新成果的基础上，对数据挖掘技术应用的各个环节做了充分的讨论。在理论部分，本书介绍了数据挖掘技术所涉及的基本概念、主流技术和最新成果，其中包含了作者的部分研究成果。在应用部分，本书结合具体的实例系统论述了商业理解、数据预处理、建模、模型部署等整个数据挖掘流程。

本书第 3 章（除 3.7 节）、第 4 章、第 5 章及第 2 章的 2.5 节由田凤占博士撰写，其余内容由胡可云博士撰写，黄厚宽教授撰写了前言，并对本书的全文进行了修改和审校。此外，陈景年、刘峰、张颜锋、王银利、冯奇同学参与了本书部分资料的整理工作，也为本书的成稿做出了贡献。

阅读本书需要读者有一定的数学分析和数理统计基础。本书既可以作为大学本科生和研究生的补充教材，也可以作为企业实施数据挖掘和商务智能的实战指导；既可以作为初次接触数据仓库和数据挖掘技术人员的入门读物，也可以作为高级研究人员的参考书。

数据挖掘及其应用内容繁多，涉及许多学科。由于作者水平有限，书中可能有不少不足之处，恳请读者不吝指正。

作　者

2008 年 3 月

目　录

理　论　篇

应 用 篇

第1章 导 论

1.1 数据挖掘概述

1.1.1 数据挖掘的背景

近年来，随着计算机技术、存储技术及互联网的快速发展，大量的数据在企业、政府部门、科研教学机构的服务器及计算机上积累起来。越来越多的企业开始建设信息化项目，在提升企业生产效率的同时，也积累了大量有关客户、生产过程、市场、经营、竞争对手、环境等方面的宝贵数据；在推行电子政务的今天，政府部门的数据也在进行数字化；科研机构更是收集了大量的科学研究数据，试图从中发现自然界和社会及经济运转的秘密。

然而，数据的迅速积累，很快超过了手工可以进行分析的规模。类似于著名的半导体摩尔定律，据统计，近年来数据内容规模以每 18 个月增长两倍的速度激增。表 1-1 给出了一些组织的大致数据量，从中可以看出当前需要处理的数据规模，而这些数据仍然在快速增长中。

表 1-1 常见的数据规模

组　织	项　目	规　模
电信公司	通话清单	数十亿条/月
银行	业务数据	数千万条/月
股票	日线数据	数十万条/月
卷烟厂	生产数据	数百万条/月
超市	销售数据	数百万条/月
环保局	污染数据	数十万条/月
搜索引擎	网页	数百亿页
电子邮件提供商	电子邮件	数千万封
在线零售商	交易数据	数百万条/月

在庞大的数据量面前，许多组织陷入了“数据丰富而信息贫乏”（Data Rich，Information Poor）的尴尬境地。这些组织投资数百万到数千万的资金在数据存储和获取各种生产、营销、研发、物流、客户数据上，获取的数据却仅仅放在那里。但是，这些数据在训练有素的数据分析挖掘人员手上却能产生巨大的回报。现在中国大多数企业没有人去分析如何应用已有的数据来进行决策，来提高客户满意程度、降低库存、优化流程等。也正因为没有充分利用信息系统中的数据进行决策分析，数据的准确、及时、完整性也没有引起足够的重视。而大量实践证明，数据的准确、及时、完整性只有在使用过程中才能不断完善和提高。将数

据转化为某种更有用的东西，需要相当多的人力投入和智慧；但大多数组织仅仅从技术的角度来看待这一问题，认为投资建立了信息系统就能解决一切问题。拥有信息系统，与拥有其他技术一样，是必要的，但对于高质量的信息和知识而言，则是不够的。信息系统相当于一个现代化的工具，然而提供竞争优势的不是工具本身，而是使用工具的方式。

随着信息系统的广泛使用，如何充分利用数据，发掘出有用的知识，是广大拥有大量数据的组织非常关心的问题。在此背景下，源自古老的数据分析及统计技术，加上现代的人工智能、数据库和统计学相关的技术，产生了数据库知识发现这门跨学科分支，人们形象地称之为“数据挖掘”——从数据中挖掘出有用的知识。

事实上在数据挖掘这个术语提出之前，统计学习和机器学习领域一直在进行类似的工作，如统计学习中的统计推断，机器学习中的规则提取方面的研究等。不同的是，数据挖掘伴随着计算机的大规模使用，所面对的问题不是在统计学习和机器学习领域中传统的小样本、定义良好的问题，而是巨大数据量、定义模糊的问题。而数据挖掘所使用的技术，主要来源仍然是机器学习和统计学领域，故数据挖掘也被称为数据库知识发现。

1995 年，第一届知识发现和数据挖掘国际会议在美国召开，标志着对知识发现和数据挖掘的研究进入了一个新的阶段。随后，数据挖掘研究在全球迅速升温，各个大学均设置了相关专业或方向。同时，数据挖掘应用也逐渐展开，目前在很多行业，如金融、电信、保险、零售、体育、制造、科学研究等，均有数据挖掘的成功应用。

随着应用的广泛，数据挖掘工具也日益流行。除了传统的数据统计分析软件商，如 SAS、SPSS，提供的数据挖掘套件外，各大数据库厂商，如 Oracle、IBM、Microsoft、NCR 等，也纷纷在其数据库产品上捆绑数据挖掘工具，以图抢占分析型数据应用的市场。

1.1.2 数据挖掘的定义

文献中数据挖掘有各种各样的定义，列举如下。

数据挖掘是识别出存在于数据库中有效的、新颖的、具有潜在效用的、最终可理解的模式的非平凡过程。

- Usama M. Fayyad 等，*Advances in Knowledge Discovery and Data Mining*；
  数据挖掘是在大型数据存储库中自动地发现有用信息的过程。
- Pang-Ning Tan 等，《数据挖掘导论（*Introduction to Data Mining*）》；
  数据挖掘就是对观测到的数据集（经常是很庞大的）进行分析，目的是发现未知的关系和以数据拥有者可以理解并对其有价值的新颖方式来总结数据。
- David Hand 等，《数据挖掘原理（*Principle of Data Mining*）》；
  数据挖掘是通过自动或半自动化的工具对大量的数据进行探索和分析的过程，其目的是发现其中有意义的模式和规律。
- Michael J. A. Barry 等，《数据挖掘——客户关系管理的科学与艺术》；

从上述各种定义可以看出，数据挖掘具有如下 3 个主要特点。

(1) 大量数据。数据挖掘所处理的数据量巨大，这也是数据挖掘方法和以前的数据分析方法不一样的一个方面。在数据挖掘算法中，算法时间复杂度很重要，否则所能处理的问题规模就会受到限制。

(2) 未知的有用的规律。数据挖掘所发现的模式或规律，应该不是显而易见的或无用

的，而是对业务来说有意义的、隐含的模式或规律。显而易见的或无用的模式如，“如某人打长途电话较多则此人电话总通话次数也较多”，发现这样的模式不是成功的数据挖掘。

(3) 是一个过程。数据挖掘不是一个神奇的黑盒子，只需要把原始数据塞进去，就能给出令人惊奇的规律。相反，它需要业务理解、数据理解、数据准备、建模、评估、部署等一系列步骤，数据挖掘人员的分析能力和业务理解能力对数据挖掘的成功扮演了重要角色。

综上所述，本书对数据挖掘的定义如下：

数据挖掘是从大规模数据集中抽取隐含的有意义的规律或模式的过程。

在实践数据挖掘过程中，有时候将抽取到的规律或模式称为数据的模型。值得注意的是，数据挖掘中有一些典型的方法，如决策树、神经网络、回归等，但是数据挖掘并不仅仅局限于这些方法，而是任何可以达到抽取数据中隐含而有意义的规律的方式。

数据挖掘不是一个完全的新学科分支。它是以统计学、机器学习、数据库等多个学科为基础的新兴学科。人们有时候会把数据挖掘和其他的一些概念相混淆，下面说明数据挖掘和一些概念的联系和区别。

1. 统计学

统计学是数据挖掘技术的主要来源之一。初学者往往不太清楚简单统计（例如，利用数据库查询语言的统计函数）和数据挖掘的区别。简单统计或查询的特点是问题的目标很明确（例如，找出数据库中有过两次不良信用记录的人有哪些），而数据挖掘问题则不那么明确，而是规律性的东西或者说是某种模式（例如，什么样的人会有不良信用记录），其挖掘结果也往往只在特定条件下才成立（例如，置信度为80%，说明所得出的结果大约在80%的时候是成立的）。而统计学的一些高级技术，如聚类、回归、判别分析、贝叶斯推断等，在数据挖掘中得到了应用。

2. 机器学习

机器学习是数据挖掘技术的主要来源之一。以前机器学习研究的领域是小规模的问题，其研究目的是为了发现机器学习的原理，强调学习算法的完备性、收敛性。而数据挖掘研究解决现实中的实际应用问题，强调挖掘过程及算法的实际可用性。机器学习是人工智能研究的一部分，数据挖掘则是多个学科技术的融合。

3. 数据仓库

数据仓库和数据挖掘这两个词经常在一起出现。但从本质上说，两者并没有太多的联系。数据挖掘并不一定必须在数据仓库上进行或者说必须先建立数据仓库才能使用数据挖掘。实际上，数据挖掘可以在任意数据集上进行。但数据仓库作为数据挖掘的数据源有一定优势，这是因为数据仓库中的数据经过了清洗、整理和聚合，在很大程度上减轻了数据挖掘数据预处理中的烦琐的数据整理负担，使得数据挖掘能迅速进入实质阶段，提高了数据挖掘的效率。此外，建立数据仓库的目的通常是进行数据展现和分析，数据挖掘是一种高级的数据分析工具，可以很好地与数据仓库一起工作。

4. 多维分析

有时候数据挖掘会和多维分析（On-Line Analytical Processing，OLAP）或者数据库统计等相混淆。多维分析和数据挖掘的目的有些相似，都是对数据进行分析，从中发现有用的模式。多维分析主要通过人工进行上滚/下钻操作，从感兴趣的角度进行察看，适合于对数据进行浅层次的了解；而数据挖掘通过对数据的各个变量之间的关系进行分析，发现数据内部

之间的关系或数据对某个特定变量（类标签）的作用，适合于发现隐藏的模式。这两者往往可以结合使用。通过多维分析发现数据中发生异常的地方（结果），再使用数据挖掘手段从中找出哪些情况下会发生这些异常（原因）。

5. 客户关系管理

在实际应用中，数据挖掘广泛应用于客户关系管理（Customer Relationship Management，CRM）领域，但这只是数据挖掘可以应用的很多领域之中的一个。一般来说，客户关系管理可以分为操作型和分析型两类。前者侧重于整个组织对客户的整体视图和规范的客户管理流程，向客户提供个性化的服务；后者则对前者提供支撑，从客户行为中通过数据挖掘手段提取客户的有关信息。常见应用如客户细分、客户流失分析、客户价值分析、交叉/向上销售等。

1.1.3 数据挖掘的应用

1. 数据挖掘典型应用

数据挖掘是一门面向应用的学科分支，在过去的几年中，已经在许多领域取得了成功的应用，特别是在计算机应用起步较早、积累了大量数据的行业，如电信、银行、零售、科学研究等行业。事实上，数据挖掘的应用范围极其广泛，限制数据挖掘应用范围的只是可用的数据和人们的想像力。下面仅列举一些目前有报道的数据挖掘的典型应用。随着数据库的广泛使用和数据挖掘技术的普及，必将有许多新颖的数据挖掘应用出现。

1）客户细分

人以类聚，客户细分或客户分群是现代营销的基础。通过聚类分析的方法，对客户进行划分，获得各个客户群不同的特征，从而对客户群进行针对性的营销，或者面向特定细分群开发特定产品，从而达到提高产品销量，提升客户忠诚度的目的。例如，银行业将客户分成不同的群体，向其提供不同的个性化投资产品。

2）客户流失预测

研究表明，保留老客户的成本远低于获取新客户的成本。但是，对所有的客户进行挽留营销不切实际并且非常昂贵。通过对客户行为模式的挖掘，客户流失预测仅找出那些可能会流失的客户。对这些客户进行针对性的挽留，可降低营销成本，提高产品收入。这对于有大量客户的电信、银行、保险等行业非常必要。

3）客户价值分析

客户对企业的贡献不同，一般来说遵循“20－80”原则，少数客户对企业的贡献占大部分比例。那么，哪些客户是企业最好的客户？仅仅是最近奉献收入最多的群体吗？哪些是潜在的好客户？通过客户价值分析，发现企业的最好客户，把有限的资源使用在能带来最大价值的客户身上。

4）异常发现

通过对数据的分析，找出其中的异常点。例如，信用卡是当今广泛使用的金融产品。随着竞争的加剧，各银行竞相大力推广信用卡，有少数不法分子趁机使用假资料申请信用卡，骗取钱财。通过数据挖掘对申请资料进行学习评分，可以发现信用欺诈的申请者，避免损失。同样可通过对洗钱模式的挖掘，发现洗钱模式以进行反洗钱；通过对税务数据的分析，发现偷税漏税行为等。

5）交叉销售/向上销售/捆绑销售

通过对商品和服务组合销售模式的分析，能够发现商品之间的搭配销售模式。利用这些模式，能够设计交叉销售或向上销售、捆绑销售策略，提升产品销售。例如，在零售业进行客户购物篮分析，根据结果对货架重新摆放，从而提高销量；电视台通过对观众观看习惯的分析，重新编排节目，提高收视率；电信公司通过客户行为分析，发现新业务的使用模式，使用捆绑销售，提高新业务的使用率；零售业巨头 Walmart 使用数据仓库和数据挖掘技术分析客户的购买模式，用于对库存的管理和销售机会的把握。

6）个性化服务

对每个人的消费模式进行分析，发现其与众不同的消费习惯，可有针对性地提供服务或进行促销。例如，在电子商务中，网站会根据过往购买记录向客户推荐新到商品；根据大多数人购买商品的行为，向客户推荐当前所购买商品的关联商品等。

7）数据库直销

一般说来，向客户随机发出的大量直销邮件，可能仅有不到 5% 的客户会做出响应。根据小规模邮件直销的结果反馈，数据挖掘能建立一个模型，找出潜在最有可能做出响应的客户，将响应率提高到 15%，从而削减了成本，提高了销量。

8）改进工作效率/过程改进

通过对日常工作/业务数据的分析，找到优化的模式，从而改进工作效率或业务流程。例如，NBA 使用了一套数据挖掘工具，分析球员的运动，以帮助教练找到最有效组织进攻和防守的方法；通过对制造厂商供应链日常活动的分析，找出供应链的最优运作方式；通过对生产计划及生产效率等数据的分析，找到最有效的排班方式；通过对生产工艺和质量数据的关系的分析，发现好的生产工艺流程等。

9）科学发现

通过对大量科学实验数据的分析，发现其中隐藏的模式，可导致新的科学发现的产生。例如，通过对天文数据的数据挖掘分析，发现新的星体；通过对生物信息数据的分析，发现新的基因和蛋白质折叠；识别具有良好药物特性的分子，以用于制造新药；通过对医疗数据的分析，发现药物和疾病之间的关系等。美国 NASA 也使用数据挖掘工具分析 2003 年哥伦比亚号失事的原因。

10）预警

通过对数据中趋势的分析，对将要可能发生的事件提出预警。例如，在电信业中，通过对以往报警数据的分析，发现有哪些常规报警可能是重大问题的前兆，并提出预警，阻止事故的发生；对工厂生产数据的分析，识别重大质量问题的前兆，以采取必要措施，避免产品质量事故的发生。

2. 对数据挖掘应用的常见误解

由于对数据挖掘应用过程的不了解，数据挖掘应用会遇到各种各样的误解。典型的有对数据挖掘存在不恰当的期望、在数据挖掘应用过程中脱离业务、使用不准确或不完整的数据进行数据挖掘而得出错误结论等。

1）不当期望

许多对数据挖掘不太了解的人具有对数据挖掘的不恰当期望。有些人对数据挖掘不屑一顾，认为它只是实验室的工具，不具有实用性。这样的认识有可能是由于不成功的数据挖掘

应用引起，也可能是因为数据挖掘工具本身具有的复杂性。还有一些人对数据挖掘期望太高，认为只要将数据置于数据挖掘工具，就能产出神奇的规则，解决疑难问题。这些不当期望都是由于不理解数据挖掘的工作方式产生的。本书致力于解释数据挖掘的理论和应用，以消除这些不当期望。

2）脱离业务

这是数据挖掘应用失败的常见原因之一。数据挖掘实施人员片面追求模型的调优而忽略从业务方面寻找合适的变量及模型的业务目标。数据挖掘应用从本质上来说是一个通过构造数据挖掘模型解决业务问题的过程，而不是实验数据挖掘技术的过程。因此，在数据挖掘过程中应根据业务目标来组织挖掘过程，围绕如何实现业务目标来设计模型。

3）数据问题

很多组织的确收集了大量的数据，但这些数据都是为了日常运营所收集，在收集过程中对于分析非常关键的数据却往往受到了冷落，这些数据在日常业务中由于收集起来比较麻烦，而在应该收集的时候被数据录入人员敷衍了事。这造成了看似有许多数据，分析却不能正常进行或者效果很差的现象。在数据收集过程中，由于收集程序方面的问题，数据中可能含有大量的错误信息，这会对数据挖掘算法产生误导，从而得出错误的结果。如果分析人员也没有加以警惕，则很有可能得出错误甚至相反的结论。例如，由于业务人员没有好好填写客户年龄信息（在业务系统中无关紧要，对数据分析较重要），在数据分析中利用了错误的数据，得出规则彩铃（一种电信业务，对方拨打电话时听见的个性化铃声）的用户大部分是老年人的错误结论。

4）唯工具论

如果在数据挖掘项目中未使用常见的如神经网络、决策树或新近流行的某种模型，则认为不属于数据挖掘。这通常是属于对数据挖掘的一种误解。在数据挖掘的定义中指出，数据挖掘是一种从大规模数据中抽取隐含的有意义的规律的过程。任何能抽取这样的规律的过程均是成功的数据挖掘，而没有必要非得使用数据挖掘中常用的模型。

1.2 数据挖掘的一般过程

1.2.1 数据挖掘中的数据集

1. 数据集的定义

数据是数据挖掘的起点。在数据挖掘中，通常把要进行分析的数据处理成一张宽表的形式，表的每一行称为一个**实例**或**对象**或**样本**，表的每一列称为**属性**或**特征**或**变量**。而且通常这张宽表在不同的挖掘算法中还被冠以不同的名称，如数据集、信息系统、样本集等。之所以对同一事物有这么多名词，这是因为各个数据挖掘方法对数据集的假设不一样。有些算法认为，数据的每一行被视为一个对象，列被视为该对象的属性或特征；另一些算法则认为，每一行均是来自某一个待处理群体的一个实例，每个实例有若干种属性或特征；还有些算法认为来源于统计学，列是一些变量（或随机变量），行是这些变量形成的分布的总体中抽取的样本。

在本书中，不再对这些术语加以区分，互相通用。

在有些数据集中，有一个特殊的属性，称为类标签。该属性指明了实例所属的类。类标

签在进行分类或聚类等数据挖掘任务时会用到。在分类时，类标签是数据挖掘学习算法的指导，数据挖掘算法根据类标签学习各类的区分规则，从而对没有类标签的新的实例进行分类（即赋予其合适的类标签值）。在聚类的时候，数据集的类标签初始为空，数据挖掘算法根据数据内在的规律给每个实例赋予合适的类标签值。

表1-2给出了一个数据集的实例。这是一个简化后的电信客户流失预测的数据集，类标签表示客户是否流失。该数据集用于数据挖掘算法学习客户流失的模式，以便用于在业务中对客户在流失前进行挽留活动。

表1-2 一个数据集实例

实例号	客户号	客户类型	年龄	月通话次数	月短信次数	…	类标签
1	1381××××××	集团客户	43	298	98	…	未流失
2	1392××××××	个人客户	32	39	26	…	未流失
3	1359××××××	测试客户	23	509	198	…	未流失
4	1390××××××	个人客户	28	190	20	…	流失
5	1591××××××	个人客户	23	59	232	…	未流失
6	1345××××××	集团客户	30	30	102	…	未流失
7	1367××××××	个人客户	49	15	498	…	未流失
8	1379××××××	个人客户	19	201	34	…	未流失
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

在实际问题中，数据挖掘所需的数据往往分布在多个来源中，需要在正式进行建模工作前将数据集中到一个数据集中，并进行缺值处理、异常点处理、变量合成和变换等工作，这一步骤称为预处理。

2. 属性数据类型

和编程语言中的变量数据类型不同，在数据挖掘过程中，数据集的属性并不是整数、浮点、布尔、字符等基本类型。这是因为数据挖掘并不强调数据之间的精确计算，而是强调发现属性或属性数据之间的关系。当然这并不是说计算不重要，相反，计算是发现数据之间关系的重要手段。

最基本的数据挖掘将属性分为如下两大类。

（1）离散型。离散型属性的特点是属性的数据之间没有确定的顺序，不能进行常规的运算。典型的离散型属性的例子，如商品的类别、电话号码、汽车品牌等。

（2）连续型。连续型属性的特点是属性的数据一般是数值，数据之间是有确定的顺序，可以互相比较和计算。典型的连续型属性的例子，如商品价格、电话拨打次数或时长、汽车的平均时速等。

值得注意的是，这两种类型的区别并不在于它们是不是用数值表示，而是在于它们本身的特征和意义。例如，电话号码也是以数值形式表示，但一般情况下，电话号码之间并不存在绝对的大小运算关系，因此它应该归类为离散型。

进一步地，根据属性数据的特征，在上述类别的基础上还可以进一步细分。

离散型可细分如下。

① 名称型。名称型是事物的名称或者称号，如在信用卡数据中的持卡人姓名、在电话数据中的电话号码。它们的作用仅仅是为了命名。这类属性所含信息量较少，在数据挖掘过程中往往被丢弃，如表 1-2 中的“客户号”。

② 类别型。类别型是一组对象的名称，如汽车品牌、信用卡的类型等。也称为集合型，如表 1-2 中的“客户类型”。

③ 序数型。序数型具有一定的顺序关系，和连续型不同的是，序数型的顺序只是个相对的顺序，数据之间的具体差距无法计算。如一场比赛的选手排名、信用卡风险级别、客户的等级等。

连续型可细分如下。

① 范围型。属性的数据在一定的范围内连续取值，如表 1-2 中的“年龄”、“月通话次数”、“月短信次数”等。

② 比例型。这类属性表示两个变量的比例。这类属性实际表达了两个属性之间关系，因而具有较多的信息量。例如，电话的每次通话时间，是总通话时长和通话次数之比。有些数据挖掘算法对能处理的属性数据类型，或者说对处理起来比较容易的属性数据类型有要求。在这种情况下，可能需要对属性数据类型进行转换。比较常见的是将连续的属性转换为离散的属性。其方法是找到连续属性中的若干个分点，使用这些分点将属性划分成若干个区间，每个区间看做一个新的属性值。此过程也称为连续属性离散化，有时候也需要对离散属性进行数值编码，使只能处理数值属性的算法可以处理这种数据。

上述属性数据类型的划分方法并不唯一，在其他文献或数据挖掘软件中可能有不同的划分方法。

1.2.2 数据挖掘的任务

数据挖掘主要有两大类主要任务：分类预测型任务和描述型任务。

分类预测型任务从已知的已分类的数据中学习模型，并对新的未知分类的数据使用该模型进行解释，得到这些数据的分类。在有些文献中，根据类标签的不同，分别称之为分类任务和预测任务。如果类标签是离散的类别，则称为分类任务；如果类标签是连续的数值，则称为预测任务。

典型的分类型任务如下：

(1) 给出一个客户的购买或消费特征，判断其是否会流失；

(2) 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；

(3) 给出一个病人的症状，判断其可能患的疾病；

(4) 给出大额资金交易的细节，判断是否有洗钱的嫌疑；

(5) 给出很多文章，判断文章的类别（如科技、体育、经济等）；

……

在分类预测任务中，数据集根据其在数据挖掘过程中扮演角色的不同，可划分为训练集、测试集、验证集。数据挖掘算法在建立模型的时候，需要用到训练集和测试集。在应用模型的时候，要用到验证集。这 3 个数据集必须来自同一数据源，并且具有类似的分布，否则数据挖掘过程会出现问题。

训练集是在数据挖掘过程中用来训练学习算法，建立模型的数据集。测试集是数据挖掘