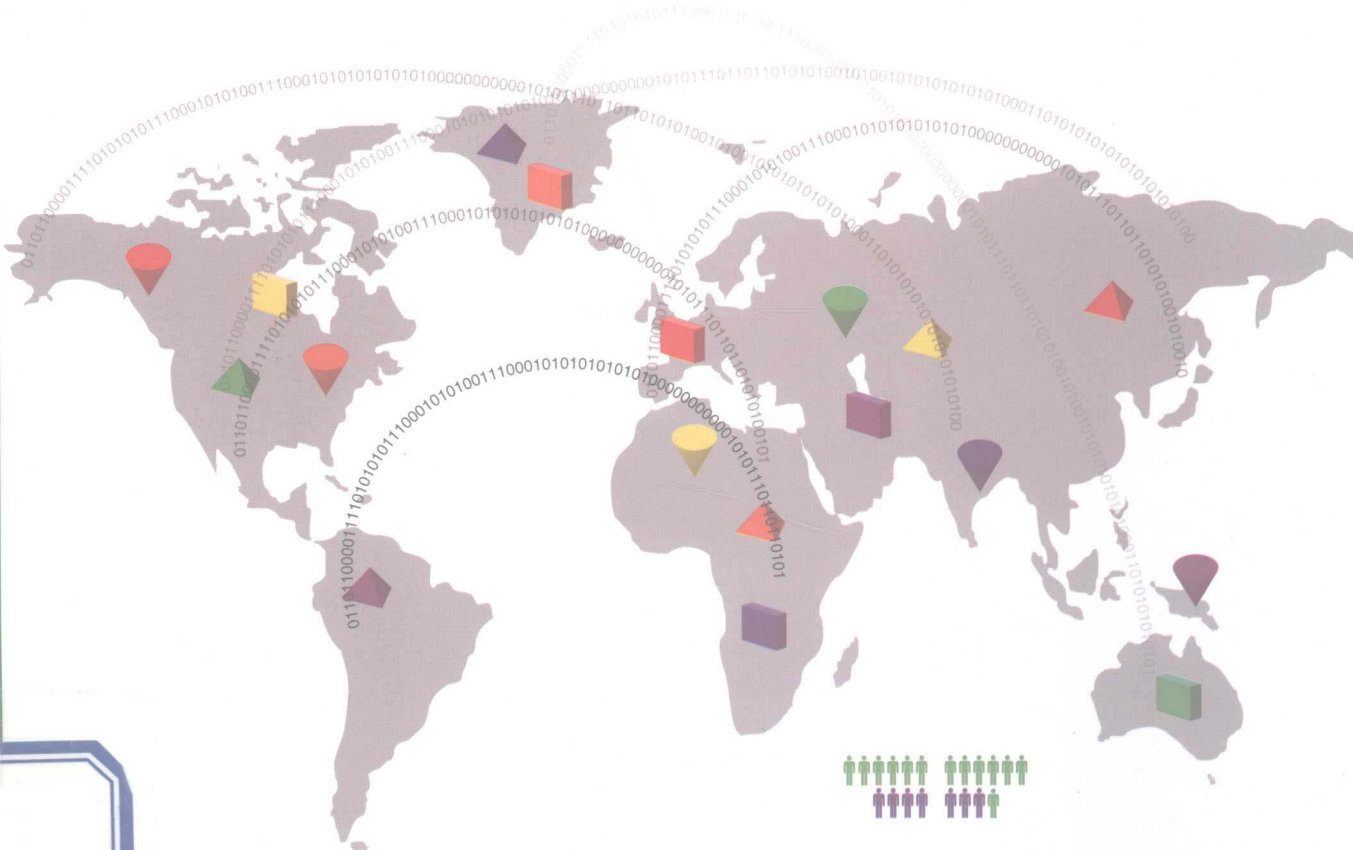
数据挖掘
理论与技术

罗森林　马 俊　潘丽敏　编著



電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

数据挖掘理论与技术

罗森林 、马 俊 潘丽敏 编著

電子工業出版社
Publishing House of Electronics Industry
北京 • BEIJING

内 容 简 介

本书梳理了数据挖掘理论与技术的知识点，注重领域内核心思想、原理、方法的论述及国内外最新研究进展的融入，内容上系统、全面、先进。全书共 9 章，主要包括数据挖掘基础知识，概率论与数理统计，数据挖掘效果评价，数据预处理，数据仓库，数据分类分析，数据聚类分析，关联规则发现，统计预测方法等。在讨论算法的同时引入应用实例，强调应用方法包括算法特点、参数选择、结果评价等方面的分析，理论联系实际，有利于算法的快速掌握和有效运用。

本书可供计算机科学与技术、生命信息工程、软件工程、通信与信息系统等相关学科、专业的学生作为教材或参考书，同时也可供科研人员参考和感兴趣者自学使用。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有，侵权必究。

图书在版编目（CIP）数据

数据挖掘理论与技术/罗森林，马俊，潘丽敏编著. —北京：电子工业出版社，2013.1

ISBN 978-7-121-18989-0

Ⅰ. ①数… Ⅱ. ①罗… ②马… ③潘… Ⅲ. ①数据采集 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2012）第 278268 号

责任编辑：曲 昕 窦 昊
印 刷：北京中新伟业印刷有限公司
装 订：北京中新伟业印刷有限公司
出版发行：电子工业出版社
北京市海淀区万寿路 173 信箱 邮编 100036
开 本：787×1 092 1/16 印张：15.75 字数：400 千字
印 次：2013 年 1 月第 1 次印刷
印 数：3 500 册 定价：46.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

前　言

数据挖掘所涉及的内容非常广泛，学科间相互交叉，互融关系复杂。本书梳理了数据挖掘理论与技术的知识点，注重领域内核心思想、原理、方法的论述及国内外最新研究进展的融入，内容上系统、全面、先进；在讨论算法的同时引入应用实例，强调应用方法包括算法特点、参数选择、结果评价等方面的分析，理论联系实际，有利于算法的快速掌握和有效运用。

本书经过长期酝酿，在总结多年的教学、应用经验的基础上认真构架而成，便于学生对数据挖掘技术的充分利用。全书共 9 章，各章的主要内容安排如下。

第 1 章绪论：包括数据挖掘的产生背景、知识基础、历史现状，技术工具、典型应用、技术难点与发展趋势等。

第 2 章概率统计理论基础：包括概率统计知识基础、随机变量的分布函数、统计推理、参数估计、假设检验、数据采样方法等。

第 3 章数据挖掘效果评价：包括模型的评分函数、模型的比较与验证、模型性能提升方法、模型的建立与理解等。

第 4 章数据预处理：包括数据预处理知识基础以及数据清理、数据集成、数据转换、数据规约、数据离散、应用实例分析等。

第 5 章数据仓库：包括数据仓库知识基础、数据仓库中的模型、数据仓库系统结构、OLAP 分析等。

第 6 章数据分类分析：主要包括分类分析的知识基础、主要技术方法及分析、贝叶斯分类、基于决策树的算法、神经网络与遗传算法、支持向量积、粗糙集与模糊集，最大熵模型、应用实例分析等。

第 7 章数据聚类分析：包括聚类分析的知识基础、主要技术方法及分析，基于划分、层次、密度、网格和模型的算法，应用实例分析等。

第 8 章关联规则发现：包括关联规则发现的知识基础、主要技术方法及分析、关联规则的基本算法、并行和分布式关联规则算法、多层次关联规则算法、数量关联规则算法，应用实例分析等。

第 9 章统计预测方法：包括统计预测的知识基础、主要技术方法及分析、回归预测方法、时间序列预测、隐马尔可夫模型、应用实例分析等。

本书由罗森林、马俊、潘丽敏共同撰写，罗森林负责整书的章节设计、内容规划和统稿，其中第 4、5、6 章由潘丽敏负责撰写，第 7、8、9 章由马俊负责撰写，其余部分由罗森林负责撰写。本书的应用实例分析主要为课题组科研成果，详细内容可查阅相关学术论文或学位论论文。

在本书的编写过程中，得到了仲顺安教授、杨煜祥老师的帮助，陈功、郭峰、郭伟东、李金玉、刘盈盈、刘峥、韩磊、王坤、韩龙飞、张蕾、王倩、陈燕颖等也做了很多工作，在此一并表示衷心的感谢。同时，衷心感谢电子工业出版社曲昕编辑对本书详细、认真的修改和热情帮助。

由于时间所限，加之编著者能力范围的限制，对于书中的不足和错误之处敬请广大读者批评指正，以便日渐完善。谢谢！

罗森林

2012 年 10 月于北京理工大学

此为试读，需要完整PDF请访问：www.ertongbook.com

目　录

此为试读，需要完整PDF请访问：www.ertongbook.com

第1章 绪 论

1.1 数据挖掘产生的背景

1.1.1 技术背景

任何技术的产生总是有它的技术背景的。数据挖掘技术的提出和普遍接受是由于计算机及其相关技术的发展为其提供了研究和应用的技术基础。纵观数据挖掘产生的技术背景，下面一些相关技术的发展起到了决定性的作用：（1）数据库、数据仓库和互联网等信息技术的发展；（2）计算机性能的提高和先进的体系结构的发展；（3）统计学和人工智能等方法在数据分析中的研究和应用。

数据库技术从20世纪80年代开始，已经得到广泛的普及和应用。在关系型数据库的研究和产品提升过程中，人们一直在探索组织大型数据库和快速访问的相关技术。高性能关系数据库引擎以及相关的分布式查询、并发控制等技术的使用，提升了数据库的应用能力。在数据的快速访问、集成与抽取等问题的解决上积累了经验。数据仓库作为一种新型的数据存储和处理手段，被数据库厂商普遍接受，相关辅助建模和管理工具快速推向市场，成为多数据源集成的一种有效的技术支撑环境。

计算机芯片技术的发展使计算机的处理和存储能力日益提高。摩尔定律告诉大家，计算机硬件的关键指标大约以每18个月翻一番的速度增长，而且现在看来仍有日益加速的趋势。随之而来的是硬盘、CPU等关键部件的价格大幅度下降，使得人们收集、存储和处理数据的能力和欲望不断提高。经过几十年的发展，计算机的体系结构，特别是并行处理技术已经逐渐成熟和普遍应用，并成为支持大型数据处理应用的基础。计算机性能的提高和先进的体系结构的发展使数据挖掘技术的研究和应用成为可能。

多年的发展中，统计学、人工智能等在内的理论与技术性成果已经成功地应用到商业处理和分析中。这些应用从某种程度上对数据挖掘技术的提出和发展起到了极大的推动作用。数据挖掘系统的核心技术和算法都离不开这些理论和技术的支持。从某种意义讲，这些理论本身发展和应用为数据挖掘提供了有价值的理论支撑和应用积累。数理统计是一个有几百年发展历史的应用数学学科，然而它和数据库技术的结合性研究应该说最近十几年才被重视。以前的基于数理统计方法的应用大多都是通过专用程序来实现的，大多数的统计分析技术是基于严格的数学理论和高超的应用技巧的，这使得一般用户很难从容地驾驭它。数据挖掘技术是数理统计分析应用的延伸和发展，假如人们利用数据库的方式从被动地查询变成主动发现知识的话，概率论和数理统计这一古老的学科可以从数据中归纳知识——数据挖掘技术提供理论基础。

人工智能是计算机科学研究中争议最多但是仍始终保持强大生命力的研究领域。机器学习应该说得到了充分的研究和发展，并且数据挖掘技术继承了机器学习解决问题的思想。专家系统（Expert System）曾经被认为是人工智能向着实用性方向发展的最有希望的技术，但是，这种技术也逐渐表现出投资大、主观性强、应用面窄等致命弱点。例如，知识获取被普遍认为是专家系统研究中的瓶颈问题。另外，由于专家系统是主观整

理知识，因此这种机制不可避免地带有偏见和错误。数据挖掘继承了专家系统的高度实用性特点，并且以数据为基本出发点，客观地挖掘知识。可以说，数据挖掘研究在继承已有的人工智能相关领域的研究成果的基础上，摆脱了以前象牙塔式的研究模式，真正开始客观地从数据集中发现蕴藏的知识。

谈到知识发现和数据挖掘，必须进一步阐述它的理论基础问题。虽然关于数据挖掘的理论基础问题仍然没有发展到完全成熟的地步，但是分析它的发展可以对数据挖掘的概念更清楚。坚实的理论是研究、开发、评价数据挖掘方法的基石。

1.1.2　理论基础

数据挖掘方法可以是基于数学理论的，也可以是非数学的；可以是演绎的，也可以是归纳的。从研究的历史看，它们是数据库、人工智能、数理统计、计算机科学以及其他方面的学者和工程技术人员，在数据挖掘的探讨性研究过程中创立的理论体系。1997 年，Mannila 对当时流行的数据挖掘的理论框架做出了综述。结合最新的研究成果，有下面一些重要的理论框架可以准确地解释数据挖掘的概念与技术特点。

1）模式发现（Pattern Discovery）架构

在这种理论框架下，数据挖掘技术被认为是从源数据集中发现知识模式的过程。这是对机器学习方法的继承和发展，是目前比较流行的数据挖掘研究与系统开发架构。按照这种架构，可以针对不同的知识模式的发现过程进行研究。目前，在关联规则（Association Rule）、分类/聚类模型（Classification/Clustering Model）、序列模式（Sequence Model）以及决策树（Decision Tree）归纳等模式发现的技术与方法上取得了丰硕的成果。

2）规则发现（Rule Discovery）架构

Agrawal 等综合机器学习与数据库技术，将三类数据挖掘目标即分类、关联和序列作为一个统一的规则发现问题来处理。他们给出了统一的挖掘模型和规则发现过程中的几个基本运算，解决了数据挖掘问题如何映射到模型和通过基本运算发现规则的问题。这种基于规则发现的数据挖掘构架也是目前数据挖掘研究的常用方法。

3）基于概率和统计理论

在这种理论框架下，数据挖掘技术被看做一个从大量源数据集中发现随机变量的概率分布情况的过程，如贝叶斯置信网络模型等。目前，这种方法在数据挖掘的分类和聚类研究和应用中取得了很好的成绩。这些技术和方法可以看做概率理论在机器学习中应用的发展和提高。统计学作为一个古老的学科，已经在数据挖掘中得到广泛应用，如传统的统计回归法在数据挖掘中的应用。统计学已经成为支撑数据仓库、数据挖掘技术的重要理论基础。实际上，大多数的理论构架都离不开统计方法的介入，统计方法在概念形成、模式匹配以及成分分析等众多方面都是基础中的基础。

4）微观经济学观点（Microeconomic View）

在这种理论框架下，数据挖掘技术被看做一个问题的优化过程。1998 年，Kleinberg 等人建立了在微观经济学框架里判断模式价值的理论体系。他们认为，如果一个知识模式对一家企业有效，它就是有趣的。有趣的模式发现是一个新的优化问题，可以根据基本的目标函数，对“被挖掘的数据”的价值提供一个特殊的算法视角，导出优化的企业决策。

5）基于数据压缩（Data Compression）理论

在这种理论框架下，数据挖掘技术被看做对数据进行压缩的过程。按照这种观点，关联规则、决策树、聚类等算法实际上都是对大型数据集的不断概念化或抽象的压缩过程。按 Chakrabarti 等人的描述，最小描述长度（Minimum Description Length，MDL）原理上可以评价一个压缩方法的优劣，即最好的压缩方法应该是概念本身的描述和把它作为预测器的最小编码长度。

6）基于归纳数据库（Inductive Database）理论

在这种理论框架下，数据挖掘技术被看做对数据库的归纳问题。一个数据挖掘系统必须具有原始数据库和模式库，数据挖掘的过程就是归纳数据查询过程。这种构架也是目前研究者和系统研制者倾向的理论框架。

7）可视化数据挖掘（Visual Data Mining）

在这种理论框架下，数据挖掘技术被看做对数据库趋势和异常的预测过程。通过应用可视化和数据挖掘技术，业务人员能够充分地探索业务数据，从而发现潜在的、以前未知的趋势，行为和异常。可视化是帮助业务人员和数据分析人员从业务数据集中发现新趋势的关键，它能够将大量复杂的模式简化成二维或三维数据集图片或数据挖掘模型。可视化数据挖掘可以认为是从数据到可视化形式再到人的感知系统的可调节的映射。可视化数据挖掘指的是采用可视化的方式检查、理解交互的数据挖掘算法。

1.1.3 数据挖掘相关概念

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型学科。人们有时候会把数据挖掘同其他一些概念相混淆，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，数据挖掘问题则不那么明确，它是规律性的东西或者说是某种模式，其挖掘结果也往往只在特定条件下才成立。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。在数据挖掘时，一种高级的数据分析工具可以很好地与数据库一起工作。

4）多维分析

有时候数据挖掘会同多维分析（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操作，从感兴趣的角度进行查看，适合对数据进行浅层次的了解；而数据挖掘通过对数据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，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。

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一。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、客户价值分析等。

1.2　数据挖掘知识基础

1.2.1　基本概念及特点

数据挖掘（Data Mining）旨在从大量的、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、人们事先不知道的而又潜在有用的信息和知识。还有很多和这一术语相近似的术语，如从数据库中发现知识（Knowledge Discovery in Database，KDD）、数据分析、数据融合（Data Fusion）以及决策支持等。

从数据挖掘的最初应用领域来看，其本质是一种新的商业信息处理技术。数据挖掘技术把人们对数据的应用，从低层次的联机查询操作，提高到决策支持、分析预测等更高级的应用上。它通过对这些数据进行微观、中观以至宏观的统计、分析、综合和推理，发现数据间的关联性、未来趋势以及一般性的概括知识等，这些知识性的信息可以用来指导高级商务活动。

从决策、分析和预测等目的看，原始数据只是未被开采的矿山，需要挖掘和提炼才能获得有用的规律性知识，这正是数据挖掘这个名字的由来。它有别于机器学习等其他研究领域，从它的提出之日起就具有很强的应用目的。数据挖掘并不是要求发现放之四海而皆准的真理，所有发现的知识都是相对的，并且对特定的行为才有指导意义。

数据挖掘与数据库中的知识发现（KDD）既有联系又有区别，从数据处理的不同层面，分析出数据挖掘自身的特点。

1）KDD 可看成数据挖掘的一个特例

既然数据挖掘系统可以在关系数据库、事务数据库、数据仓库、空间数据库（Spatial

Database)、文本数据（Text Data）以及诸如 Web 等多种数据组织形式中挖掘知识，数据库中的知识发现只是数据挖掘的一个方面。从这个意义说，数据挖掘就是从数据库、数据仓库以及其他数据存储方式中挖掘有用知识的过程。这种描述强调了数据挖掘在源数据形式上的多样性。

2）数据挖掘是 KDD 过程的一个步骤

在知识发现 1996 年国际会议上，许多学者建议对这两个名词加以区分。KDD 是从数据库中发现知识的全部过程，而数据挖掘则是此全部过程的一个特定的关键步骤。这种观点有它的合理性。虽然可以从数据仓库、Web 等源数据中挖掘知识，但是这些数据源都是和数据库技术相关的。数据仓库是由源数据库集成而来的，即使是像 Web 这样的数据源恐怕也离不开数据库技术来组织和存储抽取的信息。因此 KDD 是一个更广义的范畴，它包括数据清洗、数据集成、数据选择、数据转换、数据挖掘、模式生成及评估等一系列步骤。这样，就可以把 KDD 看做一些基本功能构件的系统化协同工作系统，而数据挖掘则是这个系统中的一个关键部分。源数据经过清洗和转换等成为适合于挖掘的数据集，数据挖掘在这种具有固定形式的数据集上完成知识的提炼，最后以合适的知识模式用于进一步分析决策。从这种狭义的观点上，可以定义数据挖掘是从特定形式的数据集中提炼知识的过程。数据挖掘作为 KDD 的一个重要步骤看待，可以更容易聚焦研究重点和有效解决问题。目前，人们对数据挖掘算法的研究基本属于这样的范畴。

3）KDD 与数据挖掘含义相同

有些人认为，KDD 与数据挖掘只是名称不一样，它们的含义基本相同。也有人说，KDD 在人工智能界更流行，数据挖掘在数据库界使用更多。但是，从广义的观点来说，数据挖掘是从大型数据集（可能是不完全的、有噪声的、不确定性的、各种存储形式的数据集）中挖掘隐含在其中，人们事先不知道的对决策有用的知识的过程。

从上面的描述中可以看出，数据挖掘概念可以在不同的技术层面上理解，但是其核心仍然是从数据中挖掘知识。数据挖掘是一个多学科交叉研究领域，融合了数据库技术、人工智能、机器学习、统计学、知识工程、面向对象方法、信息检索、高性能计算以及数据可视化等最新技术的研究成果。它不仅能对过去的数据进行查询，并且能够找出过去数据之间的潜在联系，进行更高层次的分析，以便更好地做出理想的决策、预测未来的发展趋势等。

1.2.2 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，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，每一列被视为该对象的属性或特征；另一些算法则认为，每一行被视为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；还有些算法被认为来源于统计学，列是一些变量，行是由这些变量形成的分布的总体中抽取的样本。

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。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。

一个数据集的实例如表 1-1 所示。这是一个简化后的网络宽带客户流失预测的数据集，类标签标示客户是否流失。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咨询，确认网络服务是否出现异常。

表 1-1　一个数据集实例

实例号	客户号	客户类型	年龄	月网络流量/GB	…	类标签
1	1591xxxxxxx	集团客户	33	3	…	未流失
2	1382xxxxxxx	个人客户	22	2.8	…	未流失
3	1339xxxxxxx	集团客户	33	1.6	…	未流失
4	1590xxxxxxx	个人客户	38	0.0	…	流失
5	1591xxxxxxx	个人客户	43	3.5	…	未流失
6	1515xxxxxxx	集团客户	30	2.4	…	未流失
7	1397xxxxxxx	个人客户	49	1.0	…	未流失
8	1379xxxxxxx	个人客户	29	2.6	…	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。离散性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的，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。

此为试读，需要完整PDF请访问：www.ertongbook.com