高等学校计算机专业
“十二五”规划教材

数据挖掘原理、算法及应用

李爱国　厍向阳　编著

西安电子科技大学出版社
http://www.xduph.com

内 容 简 介

本书以各类数据挖掘算法为核心，以智能数据分析技术的发展为主线，结合作者自身的研究和应用经验，阐述数据挖掘研究领域的主要理论和典型算法。全书共分 8 章：第 1 章为绪论；第 2～5 章分别介绍数据挖掘的主要技术、各类典型算法及其编程实现，包括数据预处理技术、关联规则挖掘技术、分类技术、聚类技术等几大类技术和其中包含的典型算法；第 6～8 章分别简要介绍一些数据挖掘的应用专题，包括时间序列数据挖掘、Web 挖掘、空间数据挖掘等。

本书的编写目标是让不同学术背景的研究生以及相关专业的高年级本科生理解数据挖掘技术的主要原理、各类典型算法以及这些算法的具体应用方法。

本书可作为理工科有关专业研究生和高年级本科生的教学用书，也可作为工程技术人员的参考书。

★ 本书配有部分源代码及电子教案，有需要者可从出版社网站免费下载。

图书在版编目(CIP)数据

数据挖掘原理、算法及应用/李爱国，厍向阳编著.
—西安：西安电子科技大学出版社，2012.1
高等学校计算机专业“十二五”规划教材
ISBN 978-7-5606-2731-1
Ⅰ. ① 数… Ⅱ. ① 李… ② 厍… Ⅲ. ① 数据采集—高等学校—教材 Ⅳ. ① TP274

中国版本图书馆 CIP 数据核字(2012)第 000631 号

策　　划　陈　婷
责任编辑　陈　婷
出版发行　西安电子科技大学出版社(西安市太白南路 2 号)
电　　话　(029)88242885　88201467　　邮　　编　710071
网　　址　www.xduph.com　　电子邮箱　xdupfxb001@163.com
经　　销　新华书店
印刷单位　陕西光大印务有限责任公司
版　　次　2012 年 1 月第 1 版　2012 年 1 月第 1 次印刷
开　　本　787 毫米×1092 毫米　1/16　印张 16.5
字　　数　388 千字
印　　数　1～3000 册
定　　价　29.00 元
ISBN 978-7-5606-2731-1/TP・1321
XDUP 3023001-1

前　言

数据挖掘理论和技术是20世纪80年代兴起的一门新兴交叉学科，它涉及统计学、人工智能、模式识别、机器学习以及数据库理论与技术等多门学科。数据挖掘自概念诞生以来，在学术界和工业界迅速形成了持续至今的研究和应用热潮，其地位日益重要，其应用日益广泛。随着数据库技术在工程、管理以及经济领域中的广泛应用，对数据进行后期处理和分析的需求日益广泛，而数据挖掘能够满足这种需求。因此，数据挖掘已经成为智能数据分析领域的核心技术。

本书综合当前数据挖掘领域的最新研究成果和作者本人的科学研究成果，系统地介绍数据挖掘领域的主要原理、典型算法以及应用实例。本书以各类数据挖掘算法为核心，以智能数据分析技术的发展历程为主线，结合作者自身的研究和应用经验，详细阐述数据挖掘研究领域的主要理论和典型算法及其最新进展。本书内容丰富，论述简明，力求理论联系实际，强调数据挖掘算法的分析和应用，从而使读者不仅能明白各类数据挖掘典型算法的基本原理，而且能明白如何编程实现这些算法，如何应用这些算法。

本书共分8章：第1章为绪论；第2～5章分别介绍数据挖掘的主要技术、各类典型算法及其编程实现，包括数据预处理技术、关联规则挖掘技术、分类技术、聚类技术等几大类技术和其中包含的典型算法；第6～8章分别简要介绍了一些数据挖掘的应用专题，包括时间序列数据挖掘、Web挖掘、空间数据挖掘等应用专题。

本书第1、2、6、7章由西安科技大学李爱国编写，第3、4、5、8章由西安科技大学厍向阳编写，全书由李爱国统稿。

本书的编写得到了西安科技大学研究生立项教材项目资金的资助。

作　者

2011.10

目　录

第1章 绪 论

1.1 数据挖掘的概念和定义

数据挖掘(Date Mining)是近年来随着人工智能和数据库技术的发展而出现的一门新兴技术。它是从大量的数据中筛选出有效的、可信的以及隐含信息的高级处理过程。

数据挖掘包含丰富的内涵，是一个多学科交叉的研究领域。仅从从事研究和开发的人员来说，其涉及范围之广是其他领域所难以企及的，既有大学里的专门研究人员，也有商业公司的专家和技术人员。研究背景的不同会使他们从不同的角度来看待数据挖掘的概念。因此，理解数据挖掘的概念不是简单地下个定义就能解决的问题。

1.1.1 从商业角度看数据挖掘技术

数据挖掘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。通过对特定数据进行微观、中观乃至宏观的统计、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，如顾客分析、定向营销、工作流管理、商店分布和欺诈监测等。

原始数据只是未被开采的矿山，需要挖掘和提炼才能获得对商业目的有用的规律性知识。这正是数据挖掘这个名字的由来。因此，从商业角度看，数据挖掘就是按企业的业务目标，对大量的企业数据进行深层次分析，以揭示隐藏的、未知的规律并将其模型化，从而支持商业决策活动的技术。从商业应用角度刻画数据挖掘，可以使人们更全面地了解数据挖掘的真正含义。

1.1.2 数据挖掘的技术含义

谈到数据挖掘，必须提到另外一个名词：数据库中的知识发现(Knowledge Discovery in Databases，KDD)，即将未加工的数据转换为有用信息的整个过程。KDD这个术语首次出现在1989年8月在美国底特律召开的第十一届国际人工智能联合会议的专题讨论会上。随后，在近十年的发展过程中，KDD专题讨论会逐渐发展壮大。1999年在美国圣地亚哥举行的第五届KDD国际学术大会，参加人数近千人，投稿280多篇。近年来的国际会议涉及的范围更广，如数据挖掘与知识发现(Data Mining and Knowledge Discovery，DMKD)的基础理论、新的发现算法、数据挖掘与数据仓库及OLAP的结合、可视化技术、知识表示方法、Web中的数据挖掘等。此外，IEEE、ACM、IFIS、VLDB、SIGMOD等其他学会、学刊也纷纷把DMKD列为会议议题或出版专刊，成为当前国际上的一个研究热点。

关于KDD和Data Mining的关系，有许多不同的看法。我们可以从这些不同的观点中了解数据挖掘的技术含义。

1）将 KDD 看成数据挖掘的例子之一

这一观点在数据挖掘发展的早期比较流行，并且可以在许多文献中看到这种说法。其主要观点是数据库中的知识发现仅是数据挖掘的一个方面，因为数据挖掘系统可以在关系数据库（Relational Database）、事务数据库（Transactional Database）、数据仓库（Data Warehouses）、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。从这个意义上来说，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。

2）数据挖掘是 KDD 不可缺少的一部分

为了统一认识，Fayyd、Piatetsky-Shapiro 和 Smyth 在 1996 年出版的权威论文集《知识发现与数据进展》中给出了 KDD 和数据挖掘的最新定义：KDD 是从数据中辨别有效的、新颖的、潜在有用的、最终可理解的模式的过程；数据挖掘是 KDD 中通过特定的算法在可接受的计算效率限制内生成特定模式的一个步骤。

这种观点得到了大多数学者的认同。它将 KDD 看做是一个广义的范畴，包括数据清理、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，我们可以把 KDD 看做是由一些基本功能构件组成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的部分。源数据经过清理和转换等步骤成为适合挖掘的数据集，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的分析决策工作。将数据挖掘作为 KDD 的一个重要步骤看待，可以使我们更容易聚焦研究重点，有效解决问题。目前，人们对于数据挖掘算法的研究基本属于这样的范畴。

3）KDD 与 Data Mining 的含义相同

有些人认为，KDD 与 Data Mining 只是对同一个概念的不同叫法。事实上，在现今的许多文献（如技术综述等）中，这两个术语仍然不加区分地使用着。有人说，KDD 在人工智能界更流行，而 Data Mining 在数据库界使用更多。也有人说，一般在研究领域称之为 KDD，在工程领域则称之为数据挖掘。

实际上，数据挖掘的概念有广义和狭义之分。广义的定义是，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）中，挖掘隐含在其中的、人们事先不知道的、对决策有用的知识的过程。狭义的定义是，数据挖掘是从特定形式的数据集中提炼知识的过程。

综上所述，数据挖掘概念可以从不同的技术层面上来理解，但是其核心仍然是从数据中挖掘知识。所以，有人说叫知识挖掘更合适。本书也在不同的章节使用数据挖掘的广义或狭义概念，读者要注意根据上下文加以区分。当然，在可能混淆的地方，我们将明确说明。

1.2 数据挖掘的历史及发展

数据挖掘可以看做是信息技术自然演化的结果。像其他新技术的发展历程一样，数据挖掘也必须经过概念提出、概念接受、广泛研究和探索、逐步应用和大量应用等阶段。从目前的现状看，大部分学者认为数据挖掘的研究仍然处于广泛研究和探索阶段。一方面，数据挖掘的概念已经被广泛接受；另一方面，数据挖掘的广泛应用还有待时日，需要深入

的理论研究和丰富的工程实践做积累。经过十几年的研究和实践，数据挖掘技术已经吸收了许多学科的最新成果而形成独具特色的研究。毋庸置疑，数据挖掘的研究和应用具有很大的挑战性。

随着KDD在学术界和商业界的影响越来越大，数据挖掘的研究向着更深入和实用技术两个方向发展。从事数据挖掘研究的人员主要集中在大学、研究机构，也有部分在企业和公司。所涉及的研究领域很多，主要集中在学习算法的研究、数据挖掘的实际应用以及数据挖掘理论等方面。大多数基础研究项目是由政府资助进行的，而公司的研究则更注重和实际商业问题的结合。

数据挖掘的概念从20世纪80年代被提出后，其经济价值也逐步显现出来，而且被众多商业厂家所推崇，形成初步的市场。另一方面，目前的数据挖掘系统研制也绝不是像一些商家为了宣传自己商品所说的那样神奇，而是仍有许多问题亟待研究和探索。把目前数据挖掘的研究现状描述为鸿沟(Chasm)阶段是比较准确的。所谓Chasm阶段，是说数据挖掘技术在广泛被应用之前仍有许多“鸿沟”需要跨越。例如，就目前商家推出的数据挖掘系统而言，它们都是一些通用的辅助开发工具，这些工具只能给那些熟悉数据挖掘技术的专家或高级技术人员使用，仅对应用起到加速作用，或称之为横向解决方案(Horizontal Solution)。但是，数据挖掘来自于商业应用，而商业应用又会由于领域的不同而存在很大差异。大多数学者赞成这样的观点：数据挖掘在商业上的成功不能期望于通用的辅助开发工具，而应该是数据挖掘概念与特定领域的商业逻辑相结合的纵向解决方案(Vertical Solution)。

分析目前的研究和应用现状，数据挖掘需要在如下几个方面重点开展工作。

1. 数据挖掘技术与特定商业逻辑的平滑集成问题

谈到数据挖掘和知识发现技术，人们大多引用“啤酒与尿布”的例子。事实上，目前在数据挖掘领域的确很难再找到其他类似的经典例子。数据挖掘和知识发现技术的广阔应用前景需要有效的应用实例来证明。数据挖掘与知识发现技术研究与应用的重要方向包括领域知识对行业或企业知识挖掘的约束与指导、商业逻辑有机潜入数据挖掘过程等关键课题。

2. 数据挖掘技术与特定数据存储类型的适应问题

数据的存储方式会影响数据挖掘的目标定位、具体实现机制、技术有效性等问题。指望一种能够在所有数据存储方式下发现有效知识的应用模式是不现实的。因此，针对不同的数据存储类型进行挖掘研究是目前的趋势，而且也是未来研究所必须面对的问题。

3. 大型数据的选择和规格化问题

数据挖掘技术是面向大型且动态变化的数据集的，这些数据集往往存在噪声、不确定性、信息丢失、信息冗余、数据分布稀疏等问题，挖掘前必须对数据进行预处理。另外，数据挖掘技术又是面向特定商业目标的，数据需要选择性地利用，因此，针对特定挖掘问题进行数据选择、针对特定挖掘方法进行数据规格化是数据挖掘技术无法回避的问题。

4. 数据挖掘系统的构架与交互式挖掘技术

虽然经过多年的探索，数据挖掘系统的基本构架和过程已经趋于明朗，但是在应用领域、数据类型以及知识表达模式等因素的影响下，其具体的实现机制、技术路线以及各阶段(如数据清理、知识形成、模式评估等)功能定位等方面仍需细化和深入的研究。另外，由于数据挖掘是在大量的源数据中发现潜在的、事先并不知道的知识，因此和提供源数据

的用户进行交互式探索挖掘是必然的。这种交互可能发生在数据挖掘的各个阶段，从不同角度或不同粒度进行交互。所以，良好的交互式挖掘(Interaction Mining)也是数据挖掘系统成功的前提。

5. 数据挖掘语言与系统可视化问题

对于 OLTP 应用来说，结构化查询语言 SQL 已经得到充分发展，并成为支持数据库应用的重要基石。相比 OLTP 应用而言，数据挖掘技术诞生较晚，应用更复杂，因此开发相应的数据挖掘操作语言仍然是一件极富挑战性的工作。可视化已经成为目前信息处理系统必不可少的要求，对于一个数据挖掘系统来说更是尤为重要。可视化挖掘除了要和良好的交互式技术相结合外，还必须在挖掘结果或知识模式的可视化、挖掘过程的可视化以及可视化指导用户挖掘等方面进行探索和实践。数据的可视化在某种程度上推动了人们进行知识发现，因此它可以被认为是人们从对 KDD 的神秘感变成可以直观理解知识和形象的过程。

6. 数据挖掘理论与算法研究

经过几十年的研究，数据挖掘已经在继承和发展相关基础学科(如机器学习、统计学等)方面取得了可喜的进步，并探索出了许多独具特色的理论体系。但这并不意味着挖掘理论的探索已经结束，恰恰相反，它留给研究者更多丰富的理论课题。这些研究课题一方面着眼于探索和创新面向实际应用目标的挖掘理论，另一方面的重点在于发展新的挖掘理论和算法。这些算法可能在挖掘的有效性、挖掘的精度或效率以及融合特定的应用目标等方面做出贡献。因此，对数据挖掘理论和算法的探讨将是长期而艰巨的任务。特别是，像定性定量转换、不确定性推理等一些根本性的问题还没有得到很好的解决，同时需要针对大容量数据集研究有效和高效算法。

从上面的叙述可以看出，数据挖掘研究和探索的内容是极其丰富和具有挑战性的。

1.3 数据挖掘的研究内容及功能

1.3.1 数据挖掘的研究内容

目前，数据挖掘的主要研究内容包括基础理论、发现算法、数据仓库、可视化技术、定性定量互换模型、知识表示方法、发现知识的维护和再利用、半结构化和非结构化数据中的知识发现以及网上数据挖掘。

数据挖掘所发现的知识最常见的有以下五类。

1. 广义知识(Eneralization)

广义知识指类别特征的概括性描述知识，是根据数据的微观特性发现其表征的、带有普遍性的、高层次概念的、中观或宏观的知识。反映同类事物的共同性质，是对数据的概括、精炼和抽象。

广义知识的发现方法和实现技术有很多，如数据立方体、面向属性的归约等。数据立方体还有其他一些别名，如“多维数据库”、“实现视图”、“OLAP”等。该方法的基本思想是计算某些常用的代价较高的聚集函数，诸如计数、求和、平均、最大值等，并将这些实现视

图储存在多维数据库中。既然很多聚集函数需经常重复计算，那么在多维数据立方体中存放预先计算好的结果将能保证快速响应，并可灵活地提供不同角度和不同抽象层次上的数据视图。另一种广义知识发现方法是加拿大 Simon Fraser 大学提出的面向属性的归约方法。这种方法以类 SQL 语言表示数据挖掘查询，收集数据库中的相关数据集，然后在相关数据集上应用一系列数据推广技术进行数据推广，包括属性删除、概念树提升、属性阈值控制、计数及其他聚集函数传播等。

2. 关联知识(Association)

关联知识是反映一个事件和其他事件之间依赖或关联的知识，又称依赖(Dependency)关系。这类知识可用于数据库中的归一化、查询优化等。如果两项或多项属性之间存在关联，那么其中一项的属性值就可以依据其他属性值进行预测。最为著名的关联规则发现方法是 R. Agrawal 提出的 Apriori 算法。关联规则的发现挖掘可分为两步：第一步是找出所有的频繁项集，要求频繁项集出现的频繁性不低于用户设定的最小支持度阈值(支持度反映了所发现规则的有用性)；第二步是从频繁项集中产生强关联规则，这些规则必须满足用户设定的最小置信度阈值(置信度反应了所发现规则的确定性)。识别或发现所有挖掘频繁项集是关联规则发现算法的核心，也是计算量最大的部分。

3. 分类知识(Classification&Clustering)

分类知识反映同类事物共同性质的特征型知识和不同事物之间差异的特征型知识，用于反映数据的汇聚模式或根据对象的属性区分其所属类别。最为典型的分类方法是基于决策树的分类方法。它从实例集中构造决策树，是一种有指导性的学习方法。该方法先根据训练子集(称为窗口)构造决策树。如果该树不能对所有对象进行正确的分类，那么选择一些例外加入到窗口中，重复该过程一直到形成正确的决策集。其最终结果是一棵树，叶结点是类名，中间结点是带有分枝的属性，该分枝对应该属性的某一可能值。最为典型的决策树分类系统是 ID3，它采用自顶向下不回溯策略，能保证找到一个简单的树。算法 C4.5 和 C5.0 都是 ID3 的扩展，它们将分类领域从类别属性扩展到数值型属性。

数据分类还有统计、粗糙集(Rough Set)等方法。线性回归和线性辨别分析是典型的统计模型。为降低决策树生成代价，人们还提出了一种区间分类器。最近也有人研究使用神经网络方法在数据库中进行分类和规则提取。

4. 预测型知识(Prediction)

预测型知识是指由历史的和当前的时间序列型数据去推测未来的数据，它实际上是一种以时间为关键属性的关联知识。目前，时间序列预测的经典方法有统计方法、神经网络和机器学习等。1968 年，Box 和 Jenkins 提出了一套比较完善的时间序列建模理论和分析方法，通过经典的数学方法建立随机模型，如自回归模型、自回归滑动平均模型、求和自回归滑动平均模型和季节调整模型，并在此基础上进行时间序列的预测。大量的时间序列是非平稳的，其特征参数和数据分布随着时间的推移而发生变化，仅仅通过对某段历史数据的训练，建立单一的神经网络预测模型，还无法完成准确的预测任务，为此，人们提出了统计学和基于精确性的再训练方法，当发现现存预测模型不再适用于当前数据时，对模型重新训练，获得新的权重参数，建立新的模型。此外，有许多系统借助并行算法的计算优势对时间序列进行预测。

5. 偏差型知识(Deviation)

偏差型知识是指通过分析标准类以外的特例、数据聚类外的离群值、实际观测值和系统预测值间的显著差别，对差异和极端特例进行描述。所有这些知识都可以在不同的概念层次上被发现，并随着概念层次的提升，从微观到中观、到宏观，满足不同用户不同层次决策的需要。

1.3.2 数据挖掘的功能

数据挖掘用于在指定数据挖掘任务中找到模式类型。数据挖掘任务一般可以分两类：描述和预测。描述性挖掘任务刻画数据库中数据的一般特性；预测性挖掘任务在当前数据上进行推测和预测。

用户有时不知道他们的数据中什么类型的模式是有趣的，因此数据挖掘系统要能够并行地挖掘多种类型的模式，以适应不同的用户需要或不同的应用。此外，数据挖掘系统应当能够发现各种粒度(即不同的抽象层次)的模式。数据挖掘系统应当允许用户给出提示，指导或聚焦有趣模式的搜索。由于有些模式并非对数据库中的所有数据都成立，通常每个被发现的模式需要带上一个确定性或"可信性"度量。

数据挖掘的功能主要体现在以下六个方面。

1. 类/概念描述：特征化和区分

数据可以与类或概念相关联。一个概念常常是对一个包含大量数据的数据集合总体情况的概述。对含有大量数据的数据集合进行描述性的总结并获得简明、准确的描述，这种描述就称为类/概念描述(Class/Concept Description)。这种描述可以通过下述方法得到：

(1) 数据特征化，一般地汇总所研究类(称为目标类(Arget Class))的数据。

(2) 数据区分，将目标类与一个或多个比较类(常称为对比类(Ontrasting Class))比较。

(3) 数据特征化和比较。

数据特征化(Data Characterization)是目标类数据的一般特征或特性的汇总。通常，用户指定类的数据通过数据库查询收集。例如，为研究上一年销售增加 10%的软件产品的特征，可以通过执行一个 SQL 查询收集关于这些产品的数据。

有许多有效的方法可以将数据特征化和汇总。例如，基于数据立方体的 OLAP 上卷操作可以用来执行用户控制的、沿着指定维的数据汇总。一种面向属性的归纳技术可以用来进行数据的概化和特征化，而不必一步步地与用户进行交互。

数据特征可以通过多种形式输出，包括饼图、条图、曲线、多维数据立方体和包括交叉表在内的多维表。结果描述也可以由概化关系(Generalized Relation)或规则形式(称作特征规则)提供。

数据区分(Data Discrimination)是将目标类对象的一般特性与一个或多个对比类对象的一般特性比较。目标类和对比类由用户指定，而对应的数据通过数据库查询检索。例如，用户可能希望将上一年销售增加 10%的软件产品与同一时期销售至少下降 30%的那些产品进行比较。用于数据区分的方法与用于数据特征化的方法类似。

区分描述的输出形式类似于特征描述，但区分描述应当包括比较度量，帮助区分目标类和对比类。用规则表示的区分描述称为区分规则(Discriminant Rule)。用户应当能够对

特征和区分描述的输出进行操作。

2. 关联分析

关联分析(Association Analysis)就是从给定的数据集中发现频繁出现的项集模式知识，又称为关联规则 Association Rules。关联分析广泛应用于市场营销、事务分析等领域。

通常关联规则具有 $X \Rightarrow Y$ 形式 即"$A_1 \wedge \cdots \wedge A_m \Rightarrow B_1 \wedge \cdots \wedge B_n$"的规则，其中，$A_i$ $(i \in \{1, \cdots, m\})$，$B_j (j \in \{1, \cdots, n\})$均为属性—值(属性=值)形式。关联规则 $X \Rightarrow Y$ 表示"数据库中的满足 X 中条件的记录(tuples)也一定满足 Y 中的条件"。

3. 分类和预测

分类(Classification)就是找出一组能够描述数据集合典型特征的模型(或函数)，以便能够分类识别未知数据的归属或类别(Class)，即将未知事例映射到某种离散类别之一。分类模型(或函数)可以通过分类挖掘算法从一组训练样本数据(其类别归属已知)中学习获得。

分类挖掘所获得的分类模型可以采用多种形式加以描述输出。其中主要的表示方法有：分类规则(IF-THEN)、决策树(Decision Trees)、数学公式(Mathematical Formulae)和神经网络。分类规则容易由判定树转换而成。决策树是一个类似于流程图的树结构，每个节点代表一个属性值上的测试，每个分支代表测试的一个输出，树叶代表类和类分布。神经网络在用于分类时是一组类似于神经元的处理单元，单元之间加权连接。

分类可以用来预测数据对象的类标记。然而，在某些应用中，人们可能希望预测某些空缺或未知的数据值，而不是类标记。当被预测的值是数值数据时，通常称之为预测(Prediction)。尽管预测可以涉及数据值预测和类标记预测，但预测通常是指值预测，并因此不同于分类。预测同时也包含基于可用数据的分布趋势识别。

相关分析(Relevance Analysis)可能需要在分类和预测之前进行，它试图识别对于分类和预测无用的属性。这些属性应当排除。

4. 聚类分析

聚类分析(Clustering Analysis)与分类预测方法的明显不同之处在于，后者所学习获取分类预测模型所使用的数据是已知类别属性(Class-labeled Data)，属于有监督学习方法，而聚类分析(无论是在学习还是在归类预测时)所分析处理的数据均是无(事先确定)类别归属的。类别归属标志在聚类分析处理的数据集中是不存在的。聚类也便于将观察到的内容分类编制(Taxonomy Formation)成类分层结构，把类似的事件组织在一起。

5. 孤立点分析

数据库中可能包含一些与数据的一般行为或模型不一致的数据对象。这些数据对象被称为孤立点(Outlier)。大部分数据挖掘方法将孤立点视为噪声或异常而丢弃，然而在一些应用场合，如各种商业欺诈行为的自动检测中，小概率发生的事件(数据)往往比经常发生的事件(数据)更有挖掘价值。孤立点数据分析通常称做孤立点挖掘(Outlier Mining)。

孤立点可以使用统计试验检测。它假定一个数据分布或概率模型，并使用距离进行度量，到其他聚类的距离很大的对象被视为孤立点。基于偏差的方法通过考察一群对象主要特征上的差别来识别孤立点，而不是使用统计或距离度量。

6. 演变分析

数据演变分析(Evolution Analysis)就是对随时间变化的数据对象的变化规律和趋势进行建模描述。这一建模手段包括概念描述、对比概念描述、关联分析、分类分析、时间相关数据(Time-Related)分析，时间相关数据分析又包括时序数据分析，序列或周期模式匹配，以及基于相似性的数据分析等。

1.4 数据挖掘的常用技术及工具

数据挖掘是从人工智能领域的一个分支——机器学习发展而来的，因此机器学习、模式识别、人工智能领域的常规技术，如聚类、决策树、统计等方法经过改进，大都可以应用于数据挖掘。数据挖掘的常用技术有决策树、规则发现、神经网络、贝叶斯网络、关联规则、聚类、可视化、文本/Web 挖掘等。近年来，神经网络、贝叶斯网络、关联规则等技术在数据挖掘中的应用发展很快；可视化技术受到越来越多的重视；文本和 Web 数据的挖掘成为一个新兴的研究方向。

1.4.1 数据挖掘的常用技术

数据挖掘的常用技术有：

(1) 人工神经网络：仿照生理神经网络结构的非线性预测模型，通过学习进行模式识别。

(2) 决策树：代表着决策集的树形结构。

(3) 遗传算法：基于进化理论，并采用遗传结合、遗传变异以及自然选择等设计方法的优化技术。

(4) 近邻算法：将数据集合中每一个记录进行分类的方法。

(5) 规则推导：从统计意义上对数据中的“如果—那么”规则进行寻找和推导。

采用上述技术的某些专门的分析工具已经发展了大约十年的时间，不过这些工具所能处理的数据量通常较小。现在，这些技术已经被直接集成到许多大型的符合工业标准的数据仓库和联机分析系统中了。

1.4.2 数据挖掘的工具

1. 基于神经网络的工具

神经网络用于分类、特征挖掘、预测和模式识别。人工神经网络仿真生物神经网络，本质上是一个分散型或矩阵结构，它通过训练数据的挖掘，逐步计算网络连接的加权值。由于对非线性数据具有快速建模能力，基于神经网络的数据挖掘工具现在越来越流行。其开采过程基本上是将数据聚类，然后分类计算权值。神经网络很适合分析非线性数据和含噪声数据，所以在市场数据库的分析和建模方面应用广泛。

2. 基于规则和决策树的工具

大部分数据挖掘工具采用规则发现或决策树分类技术来发现数据模式和规则，其核心是某种归纳算法。这类工具通常是对数据库的数据进行开采，产生规则和决策树，然后对新数据进行分析和预测。其主要优点是：规则和决策树都是可读的。

3. 基于模糊逻辑的工具

该方法应用模糊逻辑进行数据查询、排序等。它使用模糊概念和“最近”搜索技术的数据查询工具，可以让用户指定目标，然后对数据库进行搜索，找出接近目标的所有记录，并对结果进行评估。

4. 综合多方法的工具

不少数据挖掘工具采用了多种开采方法，这类工具一般规模较大，适用于大型数据库(包括并行数据库)。这类工具开采能力很强，但价格昂贵，并要花很长时间进行学习。

1.5 数据挖掘的应用热点

就目前来看，数据挖掘未来的几个应用热点包括网站的数据挖掘、生物数据挖掘、文本的数据挖掘、实时数据挖掘以及数据挖掘中的隐私保护和信息安全。

1. 网站的数据挖掘

随着互联网的发展，各类电子商务网站层出不穷。电子商务网站在进行数据挖掘时，所需要的数据主要来自于两个方面：一部分数据是客户的背景信息，此部分信息主要来自于客户的登记信息；另外一部分数据主要来自浏览者的点击流(Click-Stream)，此部分数据主要用于考察客户的行为表现。但有的时候，客户不肯把背景信息填写在登记表上，这就会给数据分析和挖掘带来不便。此时，就不得不从浏览者的点击流数据中来推测客户的背景信息，进而再加以分析。

就分析和建立模型的技术和算法而言，网站的数据挖掘和传统的数据挖掘差别并不是特别大，很多方法和分析思想都可以运用。所不同的是网站的数据格式有很大一部分来自于点击流，这与传统的数据库格式有区别。因而对电子商务网站进行数据挖掘所做的主要工作是数据准备。目前，有很多厂商正在致力于开发专门用于网站挖掘的软件。

2. 生物数据挖掘

生物数据具有复杂性、丰富性、重要性等特点。这些都需要在进行数据挖掘时重点关注。挖掘 DNA 和蛋白序列、挖掘高维微阵列数据、生物路径和网络分析、异构生物数据的链接分析，以及通过数据挖掘集成生物数据等都是生物数据挖掘研究的有趣课题。

生物数据挖掘和通常的数据挖掘相比，无论是数据的复杂程度、数据量还是分析和建立模型的算法，都要复杂得多。从分析算法上讲，更需要一些新的和好的算法。现在很多厂商正在致力于这方面的研究。

3. 文本的数据挖掘

随着文本数据的快速猛增，传统信息检索技术已无法满足实际的需要。文档都包含有用信息，但只有一小部分是与特定用户的需求密切相关的，在不知道文档中究竟会有哪些内容时，要想给出准确精致的查询是较为困难的。在处理大量文档时，需要对文档进行比较，评估文档的重要性和相关性，或发现多文档的模式和趋势。也可以将互联网看成是一个巨大的、动态的文本数据库。显然，随着互联网的飞速发展，文本挖掘将在数据挖掘中扮演越来越重要的角色。

4. 实时数据挖掘

许多包括流数据(比如电子商务、Web 挖掘、股票分析、入侵检测和移动数据挖掘)的应用要求能实时地建立动态数据挖掘模型。该领域还需要进一步发展。

5. 数据挖掘中的隐私保护和信息安全

Web 上有大量电子形式的个人信息，随着网上攻击能力的不断增强，对我们的隐私和数据安全造成了威胁。隐私的保护越来越得到了重视。这需要技术专家、社会科学家、法律专家和公司协作，提出隐私的严格定义和形式机制，以证明数据挖掘中的隐私保护性。

随着计算机计算能力的发展和业务复杂性的提高，数据的类型会越来越多、越来越复杂，数据挖掘将发挥出越来越大的作用。

1.6 小　结

数据库技术已经从原始的数据处理发展到开发具有查询和事务处理能力的数据库管理系统。数据库技术的进一步发展越来越需要有效的数据分析和数据理解工具。这种需求是各种应用收集的数据爆炸性增长的必然结果，这些应用包括商务和管理、生物工程、行政管理、科学和工程以及环境控制。

数据挖掘是从大量数据中发现有趣模式，这些数据可以存放在数据库、数据仓库和其他信息存储中。数据挖掘是一个年轻的跨学科领域，源于诸如数据库系统、数据仓库、统计学、机器学习、数据可视化、信息检索和高性能计算领域。其他相关领域包括神经网络、模式识别、空间数据分析、图像数据库、信号处理和许多应用领域，包括商务、经济学和生物信息学。

知识发现过程包括数据清理、数据集成、数据变换、数据挖掘、模式评估和知识表示。数据模式可以从不同类型的数据库挖掘，如关系数据库，数据仓库以及事务的、对象-关系的和面向对象的数据库。有趣的数据模式也可以从其他类型的信息存储中提取，包括空间的、时间相关的、文本的、多媒体的数据库以及万维网(www)。

数据挖掘功能包括发现类/概念描述、关联、分类和预测、聚类、孤立点分析和演变分析。特征化和区分是数据汇总的形式。

模式提供知识，如果它易于被人理解，则在某种程度上对于测试数据是有效的，并且是潜在有用的、新颖的。模式兴趣度度量，无论是客观的还是主观的，都可以用来指导发现过程。

习　题

1. 若给出一个例子，其中数据挖掘对于一种商务的成功是至关重要的，那么这种商务需要什么数据挖掘功能？它们能够由数据查询处理或简单的统计分析来实现吗？

2. 数据仓库和数据库有何不同？它们有哪些相似之处？

3. 定义下列数据挖掘功能：特征化、区分、关联、分类、预测、聚类和演变分析。

4. 区分和分类的差别是什么？特征化和聚类的差别是什么？分类和预测的差别是什么？对于每一对任务，它们有何相似之处？

第2章 数据预处理

数据库中常存在受噪声数据、空缺数据和不一致数据。现实世界的数据库十分庞大(达到TB数量级),因此如何预处理数据才能提高数据质量,提高数据挖掘结果的质量,使挖掘过程更有效、更容易成为目前研究的重点。

数据预处理的方法主要包括:数据清理、数据集成、数据变换和数据规约。

数据清理可以消除数据中的噪声,识别孤立点,纠正不一致性。数据集成将多个数据源的数据合并成一致的数据存储(如数据仓库或数据立方体)。数据变换(如规范化)将数据转换为适于挖掘的形式。数据规约可以得到数据集的规约表示,它较源数据集小得多,但仍接近于保持源数据集的完整性。这些预处理技术在数据挖掘之前使用,可以有效提高数据挖掘模式的质量,降低实际模式挖掘时的时间。

2.1 数据预处理的目的

数据源中的数据可能不完整(如某些属性值的空缺)、含噪声(具有不正确的属性值)和不一致(如同一属性的不同名称)。

不完整数据的出现可能有多种原因:某些数据被认为是不必要的,如销售事务数据中顾客的信息并非总是可用的;其他数据没有包含在内,可能只是因为输入时认为是不重要的;由于理解错误,或者因为设备故障相关数据没有记录;某些记录与其他记录的内容不一致而被删除;记录历史或修改的数据可能被忽略。空缺的数据,特别是某些属性上缺少值的元组可能需要推导。

数据含噪声可能有多种原因:数据采集设备可能出故障;在数据录入过程中发生了人为的或计算机导致的错误;可能由于技术的限制,数据传输过程中出现错误;不正确的数据也可能是由命名或所用的数据代码不一致而导致的。重复元组有时也需要进行数据清理。

数据清理(Data Cleaning)例程通过填补空缺数据平滑噪声数据,识别、删除孤立点,并纠正不一致的数据。异常数据可能使挖掘过程陷入混乱,导致不可靠的输出。

数据集成(Data Integration)指将来自不同数据源的数据合成一致的数据存储。

数据变换(Data Transformation)操作,如规格化和聚集,是将数据转换成适于挖掘的形式的预处理过程。

数据归约策略有助于从原有的庞大的数据集中获得一个精简的数据集合,并使这一精简数据集保持原有数据集的完整性。在精简数据集上进行的数据挖掘显然效率更高,并且挖掘结果与使用原有数据集的结果基本相同。概化也可以"归约"数据。概化用较高层的概念替换较低层的概念。

图2-1对上述数据预处理进行了图解。以上的数据预处理并不互斥,例如,冗余数据的删除既是数据清理,也是数据归约。