数据挖掘与聚类分析

陈 燕 李桃迎 著

SHUJU WAJUE YU JULEI FENXI

大连海事大学出版社

数据挖掘与聚类分析

陈　燕　李桃迎　著

大连海事大学出版社

图书在版编目（CIP）数据

数据挖掘与聚类分析 / 陈燕，李桃迎著 .— 大连：大连海事大学出版社，2012.11
ISBN 978-7-5632-2802-7

I. ①数… II. ①陈… ②李… III. ①数据采集 ②聚类分析 IV. ①TP274 ②O212.4

中国版本图书馆 CIP 数据核字（2012）第 284337 号

大连海事大学出版社出版

地址：大连市凌海路 1 号 邮编：116026 电话：0411-84728394 传真：0411-84727996

http:/www.dmupress.com E-mail: cbs@dmupress.com

大连住友彩色印刷有限公司印装 大连海事大学出版社发行

2012 年 11 月第 1 版 2012 年 11 月第 1 次印刷

幅面尺寸：185 mm×260 mm 印张：24

字数：545 千 印数：1～500

责任编辑：姜建军 版式设计：晓 江

封面设计：王 艳 责任校对：阮琳涵

ISBN 978-7-5632-2802-7 定价：48.00 元

本书由

中央高校基本科研业务费资助项目（2012TD019）、辽宁省教育厅科学研究一般项目（L2012173）资助出版

The published book is sponsored by

the Fundamental Research Funds for the Central Universities and

the Science Research Project of Liaoning Provincial Department of Education

作者简介

陈燕（Chen Yan），博士，大连海事大学交通运输管理学院教授/博士生导师，管理科学与工程学科一级学科带头人并为省重点学科负责人，担任辽宁省物流航运管理系统工程重点实验室主任、辽宁省创新团队负责人。曾撰写《数据挖掘技术与应用》、《数据仓库与数据挖掘》、《数据仓库技术及其应用》、《管理信息系统开发教程》、《信息经济学》等学术专著与教材。主持并完成多项国家自然科学基金、国家科技计划项目及多项省部市级项目，获得省部级奖励 10 余项，发表相关学术论文 200 余篇。

李桃迎（Li Taoying），博士，大连海事大学交通运输管理学院讲师，曾参与撰写《管理信息系统开发教程》。主持辽宁省教育厅一般项目 1 项，参与国家、省部市级项目多项，获得省部级奖励 5 项，发表相关学术论文 30 余篇。

内容提要

本书系统详细地阐述了数据挖掘及聚类分析的多种相关方法、技术及具体应用。主要内容包括：绪论，数据预处理技术，多维数据分析与组织，预测技术及应用，关联分析技术及应用，遗传算法及应用，灰色系统理论与方法，粗糙集方法及应用，基于数据挖掘的知识推理，聚类分析概述，模糊聚类，聚类融合，增量聚类，近年数据挖掘新的研究方向和文本挖掘。

本书可作为管理科学与工程、信息科学与技术、应用数学等相关专业高年级本科生和研究生的数据挖掘、知识管理、聚类分析等相关课程的教材或参考资料。同时本书有助于相关的专业研究人员提升数据挖掘与聚类分析的技巧，开拓新的研究方向。

前　言

随着互联网技术的飞速发展，全社会的信息化程度不断提高，新的管理模式不断涌现，对信息系统的依赖程度越来越高。信息管理工程研究者和管理者面临严峻挑战：如何从类型复杂、存储分散的海量数据中，迅速找出潜在有用的信息与知识？如何实现对多维数据的集中组织、分析与管理？数据挖掘可以为上述问题提供有效的解决方案。数据挖掘理论与方法的研究与创新已经成为信息科学与管理工程领域最为重要的研究方向之一。

笔者在数据仓库技术与数据挖掘模型方面潜心研究数十年。尤其近年来，通过国家教育部、科技部和交通运输部及省市多个科研项目的资助，深入研究了数据挖掘的理论、技术与方法，获得多项科研成果。特别是面向交通运输、物流管理等特色领域，开展基于数据仓库与数据挖掘的创新性研究，取得了良好的社会与经济效益。

撰写本书的目的在于：利用数据仓库技术将异构的、多维的、具有复杂特征的多源数据整合并进行集中组织与管理，在此基础上，采用多种数据挖掘方法，实现从底层信息管理到高层知识管理全过程的信息深加工、挖掘与增值。

聚类分析因其应用非常广泛而成为数据挖掘研究的重要子领域，可为探索未知的数据结构提供帮助，并能成为一系列数据分析的起点，所以聚类分析成为本书的重要内容，将聚类分析算法及其变形进行了详细介绍。本书采用逐步演算和编程运行相结合的方式，力争使广大读者通过本书的学习能够快速掌握数据挖掘模型的理论、技术与方法。

张金松、于莹莹、周琳、孙骏雄、王任远、李鹏辉等同学参与完成部分章节中具体数据挖掘与聚类分析方法的应用算例和全书的校对工作。

本书旨在涵盖典型和有代表性的数据挖掘及其中的聚类分析算法，但由于数据挖掘方法多种多样，还有许多数据挖掘模型需要进一步探讨。在编写过程中，笔者查阅了国内外大量文献资料，谨向书中提到的和参考文献中列出的学者表示感谢。如果由于我们工作的疏忽，可能本书中某处内容所参考的文献没有列出，在此向所涉及的作者深表歉意。

同时，由于时间仓促和编者能力有限，书中难免存在一些不当之处，敬请广大读者批评指正。

作　者

2012 年 8 月

目　录

第1篇　基础知识

第2篇　数据挖掘

第3篇 聚类分析

第4篇　数据挖掘新进展

随着计算机硬件与软件的高速发展，尤其是数据库技术与应用的日益普及，人们面临着快速扩张的数据海洋，如何有效利用这一数据海洋宝藏为人类服务，已成为广大信息技术工作者关注的重要焦点之一。自20世纪80年代开始，数据挖掘技术逐步开始发展起来，数据挖掘技术的迅速发展，得益于目前全世界所拥有的巨大数据资源以及对将这些数据资源转换为信息和知识资源的巨大需求，对信息和知识的需求来自各行各业，从商业管理、生产控制、市场分析到工程设计、科学探索等。数据挖掘可以视为是数据管理与分析技术的自然进化产物。

本书的第一篇将对数据挖掘的基础知识，数据预处理技术以及多维数据分析与组织进行介绍。读者通过阅读本篇，可对数据挖掘的基本知识具有初步的认识和掌握，在此基础上了解数据挖掘的应用领域以及其他与数据挖掘相关的技术。为了提高数据挖掘的质量，要预先对海量数据进行预处理，数据预处理在数据挖掘过程中也是至关重要的。第二章会对数据预处理技术进行详细的介绍，在今后的数据预处理过程中，读者可以根据实际的数据质量选择具体的数据预处理方法。在前两章的基础上，接下来将对多维数据分析与组织进行详细的介绍，读者可以在这一章中更加详细地了解到来自数据库或数据仓库中的数据特点。

第一篇的基础知识将读者带进数据挖掘这一技术的世界里，相信这部分的内容会使读者初步了解数据挖掘以及与其相关的内容，为后续知识奠定坚实的基础。

第1章 绪 论

本章简要介绍数据挖掘的概念、发展和实现数据挖掘的主要技术，对数据挖掘做出一个综合的概述，同时对数据挖掘的研究现状及应用领域进行详细阐述。在对数据挖掘概念形成总体认识后，本章将会详细分析数据挖掘和其他技术的关系，主要包括数据仓库与数据挖掘的关系，KDD与数据挖掘的关系，OLAP与数据挖掘的关系等。在进行数据挖掘时，我们通常会用到一些工具辅助实现，本章接下来将对进行数据挖掘时运用到的常见工具及评价标准做出简要介绍。另外，引出聚类的概念，并对聚类分析的研究现状进行简单综述。最后，对于本书研究内容以纲要的形式展现给读者。

本章主要是对一些概念进行描述，使读者对于数据挖掘和聚类分析有一个总体的了解，希望能在一个总体认识的环境下，对读者在后续章节的学习和阅读有更大的帮助。

1.1 数据挖掘概述

1.1.1 数据挖掘的概念

随着信息技术的发展与普及，大量的数据与信息逐渐积累，如何从海量的数据中提取有用的和有价值的信息即知识，已成为信息技术研究的重要问题，由此，数据挖掘技术应运而生。20 世纪 90 年代，以美国信息工程领域专家为代表，开始研究数据挖掘的理论与方法。

数据挖掘（Data Mining，DM）的概念最早是在 1995 年的美国计算机年会（ACM）上提出的，数据挖掘就是从大量的、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、人们事先不知道的但又是潜在有用的信息和知识的过程。

另一种比较公认的定义是 W. J. Frawley，G. Piatetsky-Shapiro 等人提出的，数据挖掘就是从存在于大型数据库的数据中提取人们感兴趣的知识。这些知识是隐含的、事先未知的、潜在的、有用的信息，提取的知识表示为概念（Concepts）、规则（Rules）、规律（Regulations）、模式（Patterns）等形式，后来专家们将这些形式的知识表达模式运用形式化定义来描述。

数据挖掘的一个重要过程就是从数据中挖掘知识的过程，也称为数据库中知识发现（Knowledge Discovery in Databases，KDD）过程和知识提取、数据采掘的过程等，并且可以在其过程中用于发现概念/类描述、分类、关联、预测、聚类、趋势分析、偏差分析和相似性分析及结果的可视化。

因此，可以将数据挖掘理解为：在庞大的数据库中寻找出有价值的隐藏事件，并利用人工智能、统计、预测的科学技术，将其数据进行科学有价值地提取和深入分析，找出其中的知识，并根据企业发展中的需求问题建立不同的挖掘模型，以此作为提供企业进行决策分析时的参考依据。

人们把原始数据看做是形成知识的源泉，就像从矿石中采矿一样。原始数据可以是结构化的，如关系型数据库中的数据，也可以是半结构化的，如文本、图形、图像数据，甚至是分布在网络上的异构数据。发现知识的方法可以是数学的，也可以是非数学的；可以是演绎的，也可以是归纳的。发现了的知识可以被用于信息管理、查询优化、决策支持、过程控制等，还可以用于数据自身的维护。数据挖掘的主要目标是：在众多复杂类型数据中找出“金块”，能在商务（企业）数据中找出提高销售量和效益的关键因素，并且也能通过数据挖掘找出影响企业效益增长的相关因素。因此，数据挖掘是一门广义的交叉学科，它汇聚了不同领域的研究者，尤其是数据库、人工智能、数理统计、可视化、并行计算等方面的学者和工程技术人员。

数据挖掘的概念随着其发展而不断得到充实，美国的一项研究报告将 DM 视为 21 世纪十大明星产业之一。数据挖掘已成为当今知识管理、商业智能领域最热门的话题之一。越来越多的企业通过对数据挖掘概念和技术的了解与应用，达到解决信息工程领域关键技术难题的目的。

1.1.2 数据挖掘的发展

第一代数据挖掘系统支持一个或少数几个数据挖掘算法。在挖掘时，挖掘算法少，数据被

一次性调入内存。这些算法被设计用来挖掘向量数据，系统的成功依赖于数据的质量。

第二代数据挖掘系统支持数据库和数据仓库。可与 DBMS 集成，或有与数据仓库相连的接口，能处理大而复杂的数据集，具有良好的可扩展性。该类系统能够挖掘大型数据集、复杂数据集和高维数据。通过支持数据挖掘模式（Data Mining Schema）和数据挖掘查询语言（Data Mining Query Language，DMQL）增加系统的灵活性，提供了与数据库和数据仓库之间的有效接口。

第三代数据挖掘系统能够挖掘 Internet/Extranet 的分布式和高度异质的数据，并且能够有效地与操作型系统集成，支持分布式和异质的数据。该类系统的关键技术之一是与预言模型无缝集成，即对建立在异质系统上的多个预言模型以及管理这些预言模型的元数据提供第一级别的支持。此外，还提供了数据挖掘系统和预言模型系统之间的有效接口。一个重要的优点是由数据挖掘系统产生的预言模型能够自动地被操作系统吸收，从而与操作型系统中的预言模块相联合，提供决策支持的功能。

第四代数据挖掘系统能够挖掘由嵌入式系统、移动系统和普遍存在的计算设备产生的各种类型的数据。目前，移动计算越来越重要，将数据挖掘和移动计算结合是当前的一个研究热点，研究开发分布式、移动式的数据挖掘系统成为第四代数据挖掘系统研究的重要课题之一。

目前，第一代数据挖掘系统仍在发展中，第二代、第三代数据挖掘系统已经出现，第四代数据挖掘系统还处于研究阶段。

现在的数据仓库存储的数据量是 GB 到 TB 级别，随着时间的推移，在未来五年，可能会达到几百个 TB 级，因此，廉价可行的存储技术对于数据挖掘来说变得非常重要。目前，普遍采用的是二级存储技术，即磁盘（磁光盘）—主存两级存储。由于缺乏快速的访问和存储磁盘技术，随着存储容量的增长、数据挖掘查询越来越复杂以及并行处理器速度的加快，存储技术可能会成为数据挖掘的新瓶颈。

根据以上阐述，给出数据挖掘的研究内容，将其研究体系归纳如图 1.1 所示：

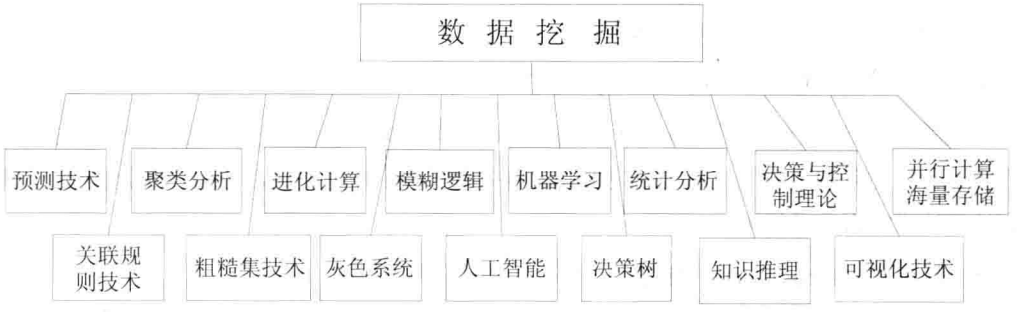

图 1.1 数据挖掘的研究体系

1.1.3 数据挖掘的主要技术

数据挖掘的主要技术包括：（1）预测技术；（2）关联规则技术；（3）聚类分析技术；（4）分类分析技术；（5）粗糙集技术；（6）进化计算技术；（7）灰色系统技术；（8）模糊逻辑技术；（9）人工智能与机器学习技术；（10）决策树技术；（11）统计分析方法；（12）知识获取、知

识表示、知识推理和知识搜索技术；（13）决策与控制理论；（14）可视化技术；（15）并行计算技术和海量存储等技术。

1.1.3.1 预测（Forecast）技术

为了科学、详细地了解某企业（某生产部门）的业务发展情况和今后的走势，可采用预测技术对其生产有利的条件进行科学论证和判断。一般在预测过程中，可以根据目标范围的不同，将其分为宏观预测和微观预测。例如宏观经济预测是指对整个国民经济或一个地区、一个部门的经济发展前景的预测；而微观经济预测是以单个经济单位的经济活动前景作为考察的对象。按预测期限长短不同，可分为长期预测，中期预测和短期预测。按预测结果的性质不同，可分为定性预测与定量预测，有的时候也采用混合预测方法。

1.1.3.2 关联规则（Association Rules）技术

数据之间的关联规则指的是在数据库中存在的一类重要的可被发现的知识。若两个或多个变量的取值之间存在某种规律性，就称为关联。关联分析的目的是找出数据库中隐藏的关联网。关联规则技术主要应用在从大型数据库中找出潜在的属性相关的知识上。例如，通过调研发现在大多数的汽车修理部门，在修理汽车的同时，也存在着购买汽车椅垫和其他零部件的可能，如果将这些相关的物品和零部件都放在汽车修理部门中，则会发现三者的效益会同时上升，从数据挖掘的角度来认识此类问题，则认为是关联知识挖掘的问题。目前，利用关联规则技术进行数据挖掘的研究非常盛行，著名的 Apriori 算法属于目前关联规则挖掘较好的算法模型之一，已经被应用在不同的研究领域中。

1.1.3.3 聚类分析（Clustering Analysis）技术

聚类分析是根据事物的特征对其进行聚类或分类，通过聚类或分类可以发现其中的规律和模式。聚类或分类以后，样本数据集就转化为类集。同一类的样本数据具有相似的变量值，不同类的样本数据的变量值不具有相似性。

1.1.3.4 分类分析（Classification Analysis）技术

分类就是找出一个类别的概念特征，用以表示这类数据的整体信息，即该类的内涵描述，进而用这种描述来构造模型，通常用规则或决策树模式表示。分类通常利用训练数据集，采用一定的算法来求得分类规则。

1.1.3.5 粗糙集（Rough Sets）技术

粗糙集技术采用的理论是粗糙集理论，将约简技术应用在不确定数据的范化和数据挖掘。粗糙集理论是波兰 Z. Pawlak 教授在 1982 年提出的一种智能决策分析理论，它是一种刻画不完整性和不确定性的数学工具，能有效地分析不精确、不一致、不完整等各种不完备的信息，并且能够将其不确定数据分析的结果即不确定和不精确的知识用已知的知识库来近似刻画和处理。利用粗糙集理论可以解决的实际问题有：不确定（不精确）数据的简化、不确定（不精确）数据的关联性发现、不确定（不精确）数据所产生的决策模型、不确定（不精确）数据所产生

的范化、基于不确定（不精确）数据的知识发现等等。目前粗糙集理论与方法已被广泛应用于不精确、不确定、不完全的信息分类和知识获取。

1.1.3.6 进化计算（Evolutionary Computation, EC）技术

基于生物界的自然选择和自然遗传机制的计算方法，如遗传算法（Genetic Algorithm，GA）、进化策略（Evolution Strategies，ES）和进化规则（Evolutionary Programming，EP）等方法，在科研和实际问题中的应用越来越广泛，并取得了较好的成果。这些方法都是基于生物进化的基本思想来设计、控制和优化人工系统，一般将这类计算方法统称为进化计算，而将相应的算法统称为“进化算法”或者“进化程序”。这些方法能够在可以接受的计算时间内，很好地解决复杂的非线性优化问题，克服具有多个局部极值的非线性最优化问题，找到全局最优解，也可以解决复杂的组合规划或者整数规划问题。

1.1.3.7 灰色系统（Grey System）技术

灰色系统是通过对原始数据的收集与整理来寻求其发展变化的规律。客观系统所表现出来的现象尽管纷繁复杂，但其发展变化有着自己的客观逻辑规律，是系统整体各功能间的协调统一。因此，如何通过散乱的数据序列去寻找其内在的发展规律就显得特别重要。灰色系统理论认为，一切灰色序列都能通过某种方式生成弱化其随机性而呈现本来的规律，认为微分方程能较准确地反映事件的客观规律，也就是通过灰色数据序列建立系统反应模型，并通过该模型预测系统的可能变化状态。

1.1.3.8 模糊逻辑（Fuzzy Logic）技术

模糊数学是继经典数学、统计数学之后，在数学上的又一新的发展。在数据挖掘领域，基于模糊逻辑可以实现模糊综合判别、模糊聚类分析等多种数据挖掘模型。

1.1.3.9 人工智能（Artificial Intelligence，AI）技术

人工智能研究计算和知识之间的关系。用机器去模拟人的智能，使机器具有类似于人的智能，其实质是研究如何构造智能机器或智能系统，以模拟、延伸、扩展人类的智能。AI 是在计算机科学、控制论、信息论、神经心理学、哲学、语言学等多种学科研究的基础上发展起来的。早期的研究领域有：专家系统、机器学习、模式识别、自然语言理解、自动定理证明、自动程序设计、机器人学、博弈、人工神经网络等；目前已涉及数据挖掘、智能决策系统、知识工程、分布式人工智能等。人工智能技术包括推理技术、搜索技术、知识表示与知识库技术、归纳技术、联想技术、分类技术、聚类技术等等，其中最基本的三种技术即知识表示、推理和搜索都在数据挖掘中得到了体现。

人工智能有许多研究领域，主要的有以下几个领域：

（1）专家系统（Expert System）。专家系统是依靠人类专家已有的知识建立起来的知识系统。目前专家系统是人工智能研究中开展较早、最活跃、成果最多的领域，广泛应用于医疗诊断、地质勘探、石油化工、军事、文化教育等各方面。它是在特定的领域内具有相应的知识和经验的程序系统，它应用人工智能技术，模拟人类专家解决问题时的思维过程来求解领域内的

各种问题，达到或接近专家的水平。

（2）机器学习（Machine Learning）。要使计算机具有知识一般有两种方法：一种是由知识工程师将有关的知识归纳、整理，并且表示为计算机可以接受、处理的方式输入计算机。另一种是使计算机本身有获得知识的能力，它可以学习人类已有的知识，并且在实践过程中总结、完善，这种方式称为机器学习。主要在以下三个方面进行机器学习的研究：一是研究人类学习的机理、人脑思维的过程；二是机器学习的方法；三是建立针对具体任务的学习系统。

（3）模式识别（Pattern Recognition）。模式识别是研究如何使机器具有感知能力，主要研究视觉模式和听觉模式的识别，如识别物体、地形、图像、字体（如签字）等。在日常生活的各方面以及军事上都有广大的用途。近年来迅速发展起来的应用模糊数学模式、人工神经网络模式的方法逐渐取代了传统的基于统计模式和结构模式的识别方法。

（4）自然语言理解。计算机如能"听懂"人的语言（如汉语、英语等），便可以直接用口语操作计算机，这将给人们带来极大的便利。计算机理解自然语言的研究有以下三个目标：一是计算机能正确理解人类的自然语言输入的信息，并能正确答复（或响应）输入的信息；二是计算机对输入的信息能产生相应的摘要，而且复述输入的内容；三是计算机能把输入的自然语言翻译成所要求的另一种语言，如将汉语译成英语或将英语译成汉语等。目前，人们做了大量的尝试，研究如何利用计算机进行文字或语言的自动翻译，但还没有找到最佳的方法，有待于更进一步深入探索。

（5）机器人学。机器人是一种能模拟人行为的机械，研究经历了三代：第一代（程序控制）机器人；第二代（自适应）机器人；第三代（智能）机器人。智能机器人具有类似于人的智能，装备了高灵敏度的传感器，具有超过一般人的视觉、听觉、嗅觉、触觉的能力，能对感知的信息进行分析，控制自己的行为，处理环境发生的变化，完成各种复杂困难的任务。而且具有自我学习、归纳、总结、提高已掌握知识的能力。目前研制的智能机器人大都只具有部分的智能，和真正意义上的智能机器人还差得很远。

（6）智能决策支持系统（IDSS）。属于管理科学的范畴，它与"知识–智能"有着极其密切的关系。将人工智能中特别是智能和知识处理技术应用于决策支持系统，扩大了决策支持系统的应用范围，提高了系统解决问题的能力，逐渐形成智能决策支持系统。

（7）人工神经网络（Artificial Neural Network）。人工神经网络从研究人脑的奥秘中得到启发，试图用大量的处理单元（人工神经元、处理元件、电子元件等）模仿人脑神经系统工程结构和工作机理。一般可分为三种网络模型：（1）前馈式网络：以感知机、误差反向传播模型、函数型网络为代表，可用于预测、模式识别等方面；（2）反馈式网络：它以 Hopfield 的离散模型和连续模型为代表，分别用于联想记忆和优化计算；（3）自组织网络：它以 ART 模型、Koholon 模型为代表，用于聚类分析等方面。

1.1.3.10 决策树（Decision Tree）技术

决策树技术主要指的是针对给定的一组样本数据，根据其对应的规则，最终选取相应的一组动作。决策树方法是利用训练集生成一个测试函数，根据不同的取值建立树的分支；在每个