“十一五”国家重点图书　　计算机科学与技术学科前沿丛书
计算机科学与技术学科研究生系列教材（中文版）

数据分析与数据挖掘

喻　梅　于　健　主　编
王建荣　王庆节　副主编

清华大学出版社

“十一五”国家重点图书　　计算机科学与技术学科前沿丛书
计算机科学与技术学科研究生系列教材（中文版）

数据分析与数据挖掘

喻　梅　于　健　主　编
王建荣　王庆节　副主编

常州大学图书
藏书章

清华大学出版社
北京

内 容 简 介

本书主要介绍数据分析与数据挖掘的基本概念和方法，包括数据的基本属性和概念、数据预处理技术、数据仓库和OLAP技术、回归分析、频繁模式挖掘、分类、聚类、离群点分析。对书中每一部分先介绍基本概念和理论基础，然后给出应用实例，便于读者更好地理解和应用算法，最后给出习题。

本书适用于数据分析与数据挖掘领域的初学者，可以作为相关专业本科及研究生教材。书中算法由浅入深、由原理到应用，有利于初学者的学习和理解。本书也可以作为数据分析与数据挖掘相关专业人士的读物。

本书封面贴有清华大学出版社防伪标签，无标签者不得销售。
版权所有，侵权必究。侵权举报电话：010-62782989　13701121933

图书在版编目(CIP)数据

数据分析与数据挖掘/喻梅，于健主编. —北京：清华大学出版社，2018
（计算机科学与技术学科前沿丛书　计算机科学与技术学科研究生系列教材：中文版）
ISBN 978-7-302-49366-2

Ⅰ. ①数…　Ⅱ. ①喻… ②于…　Ⅲ. ①数据处理－研究生－教材 ②数据采集－研究生－教材　Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2018)第014389号

责任编辑：张瑞庆
封面设计：傅瑞学
责任校对：梁　毅
责任印制：李红英

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦A座　　**邮　　编**：100084
社 总 机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn，010-62795954

印 装 者：北京密云胶印厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：18　　**字　　数**：426千字
版　　次：2018年4月第1版　　**印　　次**：2018年4月第1次印刷
印　　数：1～2000
定　　价：39.50元

产品编号：073691-01

前言

“数据分析与数据挖掘”是一门跨学科的计算机科学分支课程，是人工智能、机器学习、概率论、统计学和数据库知识的交叉学科。数据挖掘的目标是在一个或多个数据集中通过数据处理和结合一定的算法模型，最终挖掘出有价值的信息。随着科技的发展，数据量呈爆炸式增长，数据挖掘在工业界和学术界都得到了越来越多的重视。国际知名的互联网公司和科研单位都在大力发展数据科学。在我国，数据科学的发展受到了极大的关注，通过数据分析与数据挖掘可以帮助决策，进而推动经济发展。

本书内容均为数据分析与数据挖掘过程中常用方法和模型，目的是让爱好数据科学的计算机专业、统计学专业以及相关专业的学生熟悉数据挖掘的过程，掌握数据分析与数据挖掘过程中常用的算法模型及数据处理方式。本书知识点的介绍通过基础理论及概念讲解、应用例题、习题三部分进行，部分章节涉及算法应用实例。通过对数据分析与数据挖掘知识点的基础理论讲解，对数据分析与数据挖掘形成整体的认识及了解；通过应用例题，能够对算法的过程有深刻的理解；通过习题，能够巩固相应知识点。本书注重基础理论的介绍，使读者能够快速掌握数据分析与数据挖掘的基本过程和基础算法，为后续学习打下坚实的基础。

参与本书构思、撰写、审稿、应用实例的上机验证及截图校对的人员如下：喻梅、于健、王庆节、王建荣、于瑞国、陈军、邸海波、尚鸿运、孟莹、成基元、赵永伟、李鑫、曹雅茹、郭佳、刘凯、邢文涛。

本书在撰写过程中得到了清华大学出版社张瑞庆编审的大力支持，在此表示衷心感谢。

由于时间仓促、编者水平有限，书中难免有不当之处，敬请读者批评、指正。

编　者

2018 年 1 月

目录

试读结束：需要全本请在线购买：　www.ertongbook.com

第1章 概　述

本章主要介绍数据分析和数据挖掘的基本概念与基本方法，讲述对复杂、大型数据集进行分析和挖掘的重要性和必要性，简要介绍数据分析与数据挖掘的主要过程和目标，以及说明其在实际应用中存在的缺点和不足。

1.1 数据分析与数据挖掘

1.1.1 数据分析

数据分析(Data Analysis，DA)是指采用适当的统计分析方法对收集到的数据进行分析、概括和总结，对数据进行恰当的描述，并提取出有用的信息的过程。早在20世纪初期，数据分析的数学基础就已经确立，但由于数据分析涉及大量的计算，一直难以应用到实际中，计算机的出现解决了这个问题，使数据分析得到了广泛的应用。

数据分析一般具有比较明确的目标，可以根据数据分析得出的结果做出适当的判断，用来为以后的决策提供依据。例如：某连锁超市对上季度各种商品的销售量进行统计和分析，得出每种商品的需求量和销售曲线，采购部门可以根据这些数据判断是否要增加或减少订货量。

数据分析的结果可以通过列表和作图等方法表示。将数据按照一定的规律在表格中表示出来是常用的处理数据的方法，通过横向或纵向的对比可以清晰地看出数据之间的关系。表1-1为商品销售量的列表数据，可以清晰地对比四个月的销售量。

表1-1　商品销售量数据表　　(单位：件)

月份	一月	二月	三月	四月
超市一	120	118	125	122
超市二	110	115	115	120
超市三	125	120	120	125

作图法可以明确地表达各数据量之间的变化关系，常见的图有排列图、因果图、散布图、直方图、控制图等。图1-1是表1-1中数据的折线图，可以看到每个月销售量的变化情况。

1.1.2 数据挖掘

数据挖掘(Data Mining，DM)是指从海量的数据中通过相关的算法发现隐藏在数据中的规律和知识的过程。

试读结束：需要全本请在线购买：　www.ertongbook.com

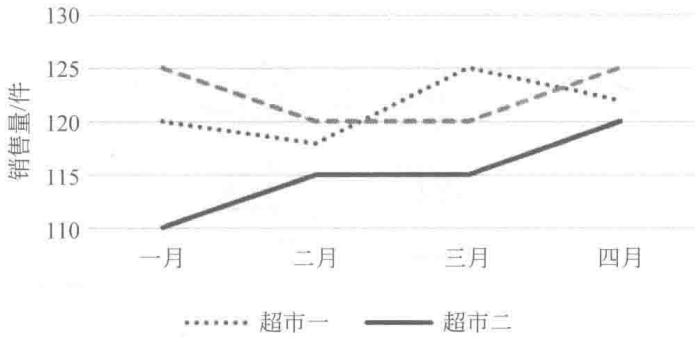

图 1-1 商品销售量数据图

实际上,“数据挖掘”一词并不能完全地表达其含义,更准确的表达应当是“在大量数据中挖掘知识”,数据挖掘又称为“资料勘探”“数据采矿”,类似于在大量的沙子中挖掘金矿,数据挖掘强调在大量的、未经过加工的数据中发现少量的、具有重要价值的知识。

在计算机行业中,数据挖掘是发展最快的领域之一。随着计算机技术的飞速发展和迅速普及,一个不得不面临的问题就是每时每刻都在产生大量的数据。例如:在线交易网站每天成交上千万的订单;哈勃望远镜每周产生约 120GB 的观测数据;某即时交流工具有数亿人同时在线;医疗行业每天有大量的诊疗病历产生。科研机构和企业投入了大量的人力和物力收集和保存这些数据,然而只有其中一小部分的数据能够被充分利用。由于数据量巨大、数据结构复杂,在很多情况下无法进行有效分析。因此,如何对这些数据进行处理并发现具有重要意义的知识是一个非常严峻的问题。

通常将数据挖掘视为数据中“知识发现”的同义词,也可以认为数据挖掘是知识发现中的一个步骤。知识发现的过程如下。

① 数据清理:消除数据中的噪声。

② 数据集成:将不同来源的数据组合在一起。

③ 数据选择:从数据库中选择与任务相关的数据。

④ 数据变换:将数据变换成适合挖掘的形式。

⑤ 数据挖掘:使用数据挖掘的方法发现知识。

⑥ 模式评估:识别知识中有用的模式。

⑦ 知识表示:将挖掘到的知识用可视化的技术表示出来。

知识发现过程如图 1-2 所示。

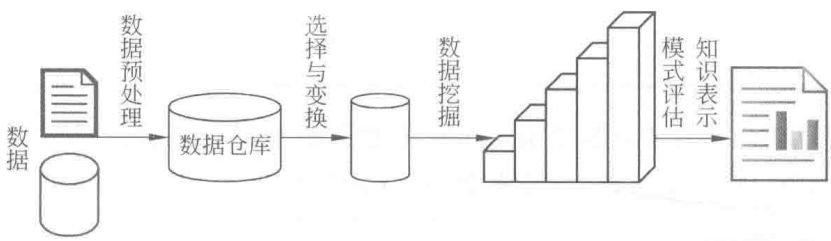

图 1-2 知识发现过程

图 1-2 中的“数据预处理”包括“数据清理”和“数据集成”两个步骤。

当提到“数据挖掘”时,通常情况下要表述的是知识发现的整个过程。因此,本书提到的

数据挖掘也是其广义的含义。

1.1.3　区别和联系

由数据分析与数据挖掘的定义可知，二者具有如下区别。

① 数据挖掘处理的是海量的数据，这里用了“海量”而不是“大量”，表示数据挖掘处理的数据量极大；而数据分析处理的数据量不一定很大。

② 数据分析往往有比较明确的目标，而数据挖掘所发现的知识往往是未知的，需要通过数据挖掘的方法发现隐藏在数据中的有价值的信息和知识。

③ 数据分析着重于展现数据之间的关系，而数据挖掘可以通过现有数据并结合数学模型，对未知的情况进行预测和估计。

下面的例子说明了数据分析与数据挖掘的不同之处。

在将要举办的生日聚会中，只有 300 元的预算，为了将聚会举办得更加体面，组织人花费了一个下午的时间调查了肉类、蔬菜、水果、饮料以及生日蛋糕的价格，经过整理和分析得到一张表格，内容是每个店铺中各种食材的价格，以便对比和选择，这个过程称为数据分析。但显然不能因为白菜的价格低而举办一场“白菜盛宴”，因此，应该考虑到好友的口味、各种食材的营养价值、食材之间的搭配以及做饭和用餐的时间，最后综合考虑这些信息，得出一个最有性价比的采购方案，使这场聚会更加完美，这个过程称为数据挖掘。

然而，数据分析与数据挖掘又联系紧密、相辅相成，数据分析的结果往往需要进一步地挖掘才能得到更加清晰的结果，而数据挖掘发现知识的过程也需要对先验约束进行一定的调整而再次进行数据分析。数据分析可以将数据变成信息，而数据挖掘将信息变成知识。如果需要从数据中发现知识，往往需要数据分析和数据挖掘相互配合，共同完成任务。

1.2　分析与挖掘的数据类型

数据分析与数据挖掘是一种通用的技术，可以应用于各种不同类型的数据，只要数据中包含一定的实际价值，都应当可以被分析和挖掘。数据的常见形式有数据库数据、数据仓库数据、事务数据等，本节将对这些数据类型进行简单介绍。

1. 数据库数据

数据库系统(DataBase System，DBS)是由一组内部相关的数据(称为数据库)和用于管理这些数据的程序组成，通过软件程序对数据进行高效的存储和管理并发、共享或分布式访问，当系统发生故障时，数据库系统应当保证数据的完整性和安全性。

关系数据库是目前使用较为成熟的数据库形式，基于关系数据库模型的数据库是数据表的集合，其中每个表都有一个唯一的名字。每个表格包含一个或多个用列表示的数据属性，每行包含一个数据实体，被唯一的关键字标识，并被一组属性描述。在创建数据表时，可以根据某列属性值的数据范围进行进一步的约束，例如标识员工年龄的列不可能出现小于 0 的值，当然，出现很大的值(如 1000)也是不合理的。

例如，某超市的商品销售情况可以用关系数据表表示，如表 1-2 所示。

表 1-2 商品销售记录

商品编号	商品名称	商品单价(元/kg)	购买数量(kg)	总价(元)
100001	苹果	6	2	12
100002	香蕉	5	3	15
100003	鸭梨	3	4	12
…	…	…	…	…

实际上,用于存储商品销售记录的表还会包含很多数据,例如每个顾客会购买多种商品,某个顾客的一次购物数据组成一个订单,数据库需要记录购物的时间、应收取金额、实际收取金额等数据,有时超市会进行促销,商品的折扣率、折扣产生的金额也应当详细地记录在数据表中。

关系数据库中的数据可以通过数据库查询进行访问,数据库查询使用关系查询语言,如SQL(Structured Query Language,结构化查询语言)。一个给定的查询语句通过数据库软件程序的处理被转换成一系列关系操作,如连接、选择、投影等。例如,可以通过关系查询获得"三月份苹果的销售量是多少""本季度哪种商品销售量最高"或"哪个月的总收入最高"等数据。

当对关系数据库进行数据挖掘时,可以通过进一步的分析和挖掘发现更有意义的模式,例如,不同年龄段的顾客对商品的喜好程度、哪些商品的销售量与月份相关、哪些商品通常会同时出现在一张订单中以及商品包装和口味的变化对销售量有什么影响等。通常来说,这些问题是商家更加关注的。

2. 数据仓库数据

假设上面提到的超市是一个连锁超市,它在全国有许多连锁店,由于销售水平和面向群体的不同,不同的区域需要单独管理数据库,当需要对所有的数据进行分析的时候,可能就会面临数据分散等问题,这时就需要用到数据仓库(Data Warehouse,DW)。

数据仓库使用特有的资料存储架构,对数据进行系统的分析整理。数据仓库通过数据清理、数据变换、数据集成、数据装入和定期数据刷新构造,本书第 4 章将详细介绍数据仓库的相关知识。图 1-3 描述了数据仓库构造和使用的过程。

数据库的数据组织是面向任务的,而数据仓库中的数据则是按照主题进行组织的。主题是指决策者进行决策时所关心的重点内容。例如:连锁超市的总经理不会关心某个超市每天卖出了几个苹果,他关心的是每个地区、每种商品的销售数据的汇总。此时,商品销售即为主题。

通常,数据仓库使用数据立方体的多维数据结构建模,其中每个维度包含模式中的一个或一组属性,而每个单元保存对应的属性值。数据立方体可以从多个维度观察数据,为决策者提供整体的信息。

联机分析处理(On-Line Analysis Processing,OLAP)是数据仓库系统的主要应用,用于支持复杂的分析操作,允许在不同的汇总级别对数据进行汇总。

数据仓库对数据的分析提供了强大的支持,但进行更加深入的分析依然需要数据挖掘工具的帮助。关于数据仓库、联机分析处理技术等将在第 4 章进行更加详细的介绍。

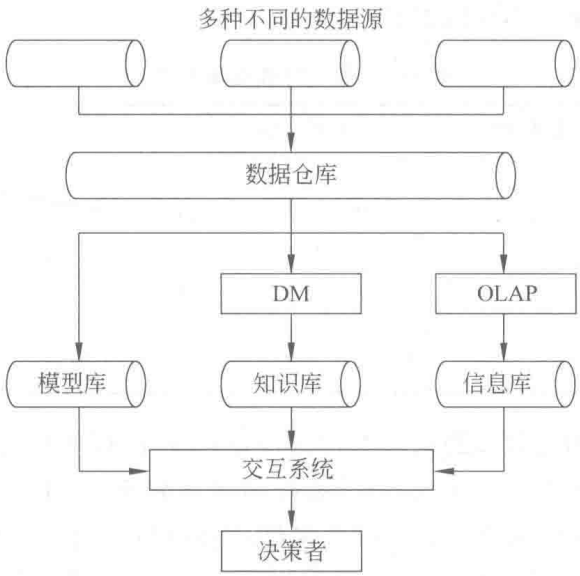

图 1-3　数据仓库的构造和使用过程

3. 事务数据

事务数据库的每个记录代表一个事务，例如一个车次的订票、顾客的一个订单等。通常来说，一个事务由一个唯一的标识号和一组描述事务的项组成，有时也需要一些附加信息表示事务的其他信息，如对商品的描述等。

依然以超市销售的商品为例，一个商品销售的事务如表 1-3 所示。

表 1-3　销售事务数据表

事务编号	商品编号
T1001	1,2,5,7,12
T1002	2,5,8,10
…	…

通过这样的数据表，可以发现多个项在一个事务中同时出现，这在现实中有重要的意义。例如：购买了牛奶的顾客很可能会同时购买面包。通过这些事务数据，决策者可以做出相应的促销策略，如将面包和牛奶放置在相近的位置，以期销售更多的商品。

4. 数据矩阵

在一个数据集中，如果数据对象的所有属性都是具有相同性质的数值型数据，那么这个数据集就可以用矩阵表示。例如表 1-4 鸢尾花数据集的部分数据实例。该数据集由三种不同类型的鸢尾花组成，其中每种类型具有 50 个样本。表中每一行代表一个数据对象，可以看作是多维空间中的一个点，每一列代表数据对象的一种属性，m 个数据对象和 n 个属性构成一个 $m \times n$ 的数据矩阵。

通过将每个数据对象映射到多维空间中的点(或向量)，可以根据数据对象的空间位置关系进行分类和聚类操作，空间上距离相近的两个数据对象被认为是同一个类型，而空间上

距离较远的两个数据对象是不同的类型。

表 1-4 鸢尾花数据集实例 （单位：cm）

	花萼长度	花萼宽度	花瓣长度	花瓣宽度
Setosa	5.1	3.5	1.4	0.2
Setosa	4.9	3.0	1.4	0.2
Versicolor	7.0	3.2	4.7	1.4
Versicolor	6.4	3.2	4.5	1.5
Virginica	6.3	3.3	6.0	2.5
Virginica	5.8	2.7	5.1	1.9

一个更加典型的应用是对文档的分类，根据不同文档出现某些关键词的频率的不同，可以将文档划分为不同的类型。一个文档-词矩阵的实例如表 1-5 所示，其中，表格中的数据为关键词出现的次数，由于频数有较大的局限性，实际应用中会采用更加具有代表性的方法，如 TF-IDF 方法。根据表中的数据可以看出，文档 1 和文档 2 具有相同的类型，文档 3 和文档 4 具有相同的类型，如果考虑到词的具体含义，则前两个文档偏向于介绍数据挖掘，而后两个文档介绍的很可能是算法。

表 1-5 文档-词矩阵实例

	数据挖掘	数据分析	算　法	复杂度
文档 1	4	3	2	1
文档 2	4	4	1	1
文档 3	0	1	6	3
文档 4	0	0	7	3

5. 图和网状数据

图和网状结构通常用来表示不同节点之间的联系，如人际关系网、网站之间的相互链接关系等。例如，通过分析微博上的人脉关系可以得到不同群体的喜好，以及哪些人被关注的程度很高，对热点话题起主导作用。图和网状数据往往包含重要的信息，但其结构复杂，对数据分析和数据挖掘提出了较高的要求。

一个典型的应用就是搜索引擎对网站页面链接关系的分析。一般来说，被指向次数越多的网页，其重要程度越高，被指向次数较少的网页，其重要程度较低。搜索引擎通过分析海量的网页链接关系，找出重要程度更高的网页反馈给用户，得到更好的搜索结果。如图 1-4 所示，每个节点代表一个网页，有向边表示网页间的链接关系。著名的网页排名算法 PageRank 就是通过分析网页之间的链接关系给出网页的重要程度的。

6. 其他类型的数据

除了上述提到的关系数据库数据、数据仓库数据、事务数据、数据矩阵以及图和网状数据以外，还有许多不同形式的其他数据，如与时间相关的序列数据（不同时刻的气温、股票市场的历史交易数据等）、数据流（监控中的视频数据流等）、多媒体数据（视频、音频、文本和图像数据等）。这些不同形式和结构的数据给数据分析和数据挖掘带来了新的挑战。

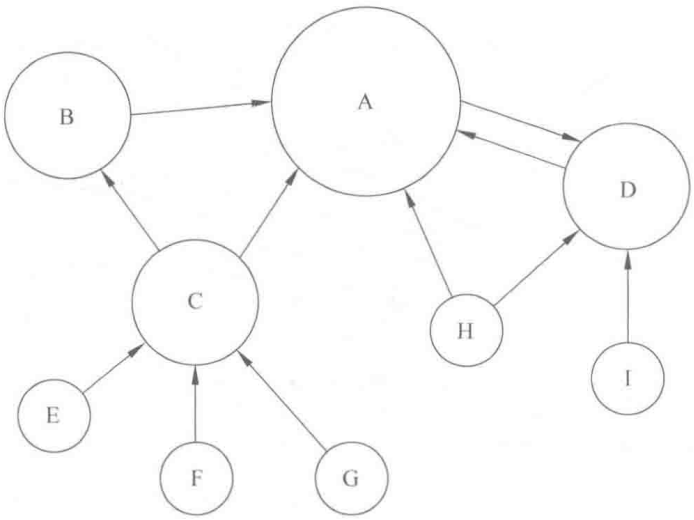

图 1-4　网页链接关系图

这些类型的数据中也包含着各种知识。例如，可以通过挖掘股票市场的历史交易数据发现股票的趋势，制定合理的投资策略；通过挖掘地铁站不同时间段的客流量数据，并根据挖掘结果安排列车的首末班时间以及两辆列车之间的时间间隔；通过挖掘不同时间段车辆流量信息，并根据得到的结果调整交通指示灯的时间，达到更高的运输效率；通过挖掘与“数据挖掘”领域相关的文献，可以了解该领域在不同的历史时期关注的热点问题的演变；通过挖掘在线销售平台上顾客发表的评论，可以根据不同顾客的意见提供更好的服务。分析和挖掘这些类型的数据可能需要更复杂的机制，但它们也为数据挖掘提出了具有挑战性和现实意义的问题。

1.3　数据分析与数据挖掘的方法

1. 频繁模式

顾名思义，频繁模式就是在数据集中频繁出现的模式。通常来说，多次出现的事物可能具有特殊的意义。因此，挖掘频繁模式可以发现包含在数据集中的有趣的关联。

频繁模式广泛应用于信用卡分析、患者就诊分析以及购物车分析等方面，其中，购物车分析在生活中最为普遍。在超市中，如果知道哪些商品经常一起出售，就可以将这些商品摆放在距离较近的位置，既方便了顾客选购，又能增加销售量。

2. 分类与回归

分类是指根据已经具有类别标签的数据集建立分类模型，并通过该模型预测不具有类别标签的数据属于哪种类别。常见的分类算法有决策树、朴素贝叶斯分类、支持向量机以及神经网络等。一个神经网络的示意图如图 1-5 所示，输入层为身高和体重的数据，中间层为输入数据在高维空间中的特征表示，输出层表示回归的结果，数值较大的维度对应的结果为预测值。

分类是通过建立模型预测离散的标签(类别)，而回归则是通过建立连续值模型推断新

试读结束：需要全本请在线购买：　www.ertongbook.com