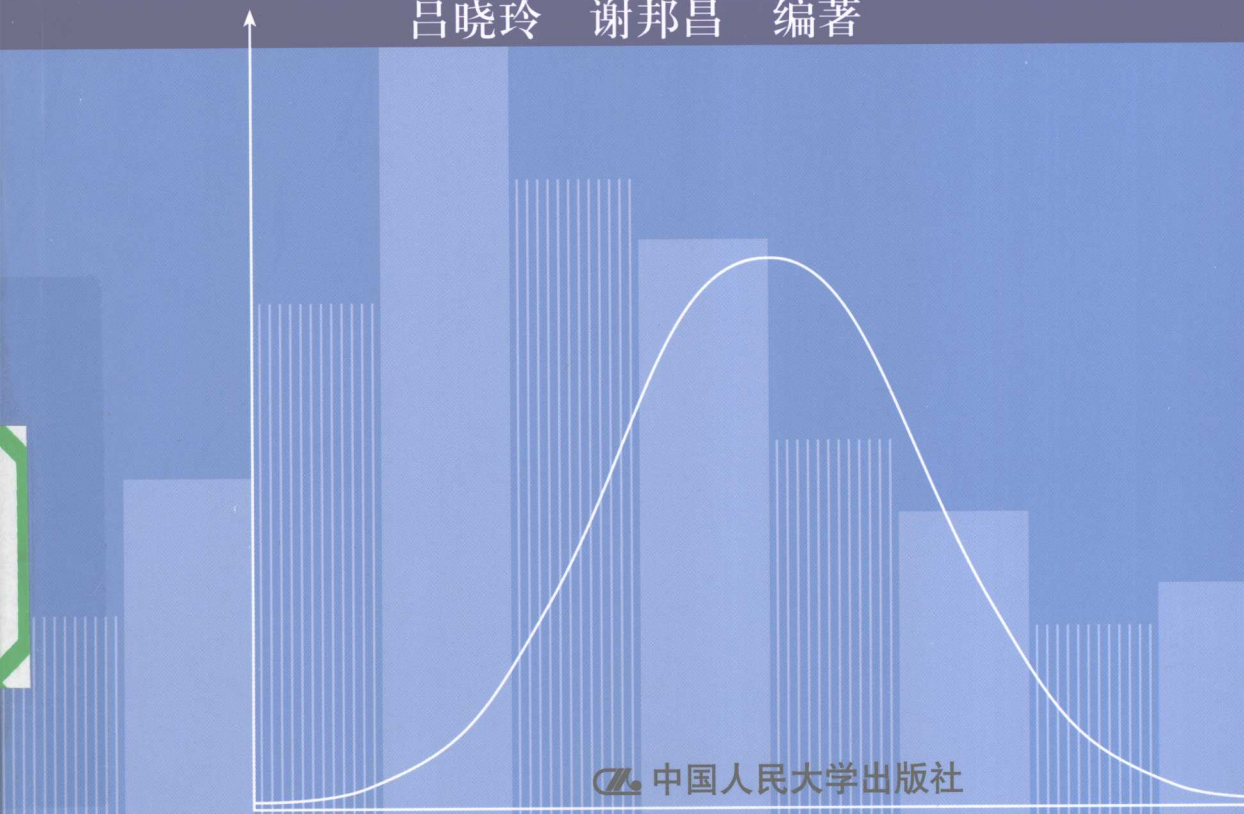
中国人民大学统计咨询研究中心
中国人民大学数据挖掘中心
中国人民大学概率论与数理统计研究所
教育部重点科研基地应用统计科学研究中心

联合推出

数据分析系列教材

数据挖掘方法与应用

吕晓玲　谢邦昌　编著



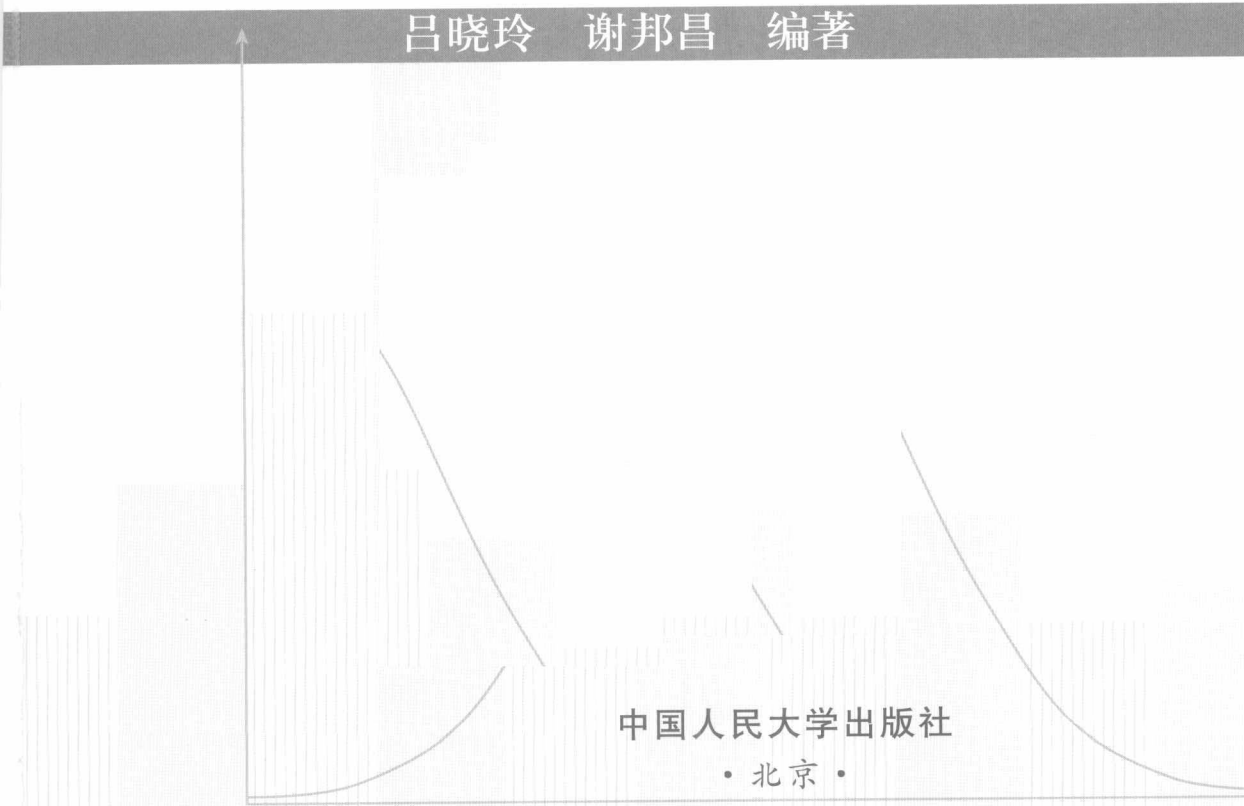
中国人民大学出版社

中国人民大学统计咨询研究中心
中国人民大学数据挖掘中心
中国人民大学概率论与数理统计研究所
教育部重点科研基地应用统计科学研究中心
联合推出

数据分析系列教材

数据挖掘方法与应用

吕晓玲　谢邦昌　编著



中国人民大学出版社
·北京·

图书在版编目（CIP）数据

数据挖掘：方法与应用/吕晓玲，谢邦昌编著.
北京：中国人民大学出版社，2008
（数据分析系列教材）
ISBN 978-7-300-09970-5

Ⅰ. 数…
Ⅱ. ①吕…②谢…
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 177676 号

数据分析系列教材
数据挖掘：方法与应用
吕晓玲　谢邦昌　编著

出版发行	中国人民大学出版社		
社　　址	北京中关村大街 31 号	**邮政编码**	100080
电　　话	010－62511242（总编室）		010－62511398（质管部）
	010－82501766（邮购部）		010－62514148（门市部）
	010－62515195（发行公司）		010－62515275（盗版举报）
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com(人大教研网)		
经　　销	新华书店		
印　　刷	北京丰印诚印务有限公司		
规　　格	170 mm×228 mm　16 开本	**版　　次**	2009 年 1 月第 1 版
印　　张	15.75 插页 1	**印　　次**	2009 年 1 月第 1 次印刷
字　　数	278 000	**定　　价**	23.00 元

总　序

随着社会经济的不断发展、科学技术的不断进步，统计方法越来越成为人们必不可少的工具和手段。在教学过程中，老师们也越来越感到运用统计方法解决实际问题的重要，不少人在探索如何运用统计软件介绍和学习统计方法。谢邦昌教授、黄登源教授在多年的教学中，积累了丰富的经验，他们热情倡议，将他们的讲稿提供出来并编写成教材，供更多的人学习和使用。这正与我们的初衷不谋而合。2005 年开始着手这套系列教材的编写，经过不断讨论、反复的论证，形成了现在的模式。由于有许多研究生的帮忙，又有几位年轻老师的辛劳，这套书终于问世。

在我们看来，掌握统计方法不仅要理论上弄明白，更重要的在于能够正确有效地运用这些方法，分析说明实际问题。这套书正是试图利用实际数据，通过统计软件的实际操作，将所能够使用的统计方法加以说明，使读者不仅能够了解相应的统计方法，而且能够通过计算机操作学会运用这些方法处理分析实际数据。希望本套书的出版能够为读者提供这样学习的工具。

由于水平有限，难免有不足之处。恳请读者朋友们提出宝贵意见。我们也会循着这样的思路，在教学以及和读者的交流沟通中不断积累、不断提高、不断完善，奉献给读者更多更好的成果。

感谢为这套书的编写付出汗水的研究生，感谢几位认真用心的年轻老师，感谢中国人民大学出版社的大力支持。为方便读者，书中的所有例题数据，都将放在中国人民大学出版社工商管理分社的网站（www. rdjg. com. cn）上，供读者下载并练习。谢谢读者，希望能够加强沟通和联系，为提高统计方法实际运用的能力和水平共同努力。

易丹辉

前　言

随着信息技术的飞速发展，数据的产生和存储达到了空前繁荣的阶段。如何从海量的数据中提取潜在的有用信息，给传统的数据处理技术提出了严峻的考验，数据挖掘方法应运而生。数据挖掘是一个多学科的交叉研究领域，不仅大学里的学术人员在研究它，商业公司的专家和技术人员也在密切地关注它和使用它；它不仅涉及人工智能领域以及统计学的应用，而且涉及数据库的管理和使用。从技术上来讲，数据挖掘是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的，但又潜在有用的信息和知识的过程。从商业应用来讲，数据挖掘是一种新的商业信息处理技术，其主要特点是对商业数据库中的大量业务数据进行抽取、转换、分析和其他模式化的处理，从中提取辅助商业决策的关键性数据。

本书第 1 章对数据挖掘进行了概述，包括数据挖掘的定义、重要性、功能、步骤和常用方法等。第 2 章和第 3 章介绍了两种数据挖掘中常用的无指导的学习算法、关联规则和聚类分析，它们处理的数据的特点是没有独立的需要预测或分类的变量，而只是试图从数据中发现一些固有的模式。关联规则就是要发现两个或多个事物之间的联系；聚类分析就是要把数据中具有相似性质的放在一类，而不同类之间尽量做到有较大的不同。第 4 章和第 5 章介绍了两种数据挖掘中常用的指导的学习算法、决策树和神经网络。它们处理的数据含有独立的需要预测或分类的变量，它们的目的就是寻找一些自变量的函数或算法对数据进行准确的预测或分类。决策树方法在对数据处理的过程中，将数据按照树状结构分成若干分

枝形成决策规则；神经网络在一定程度上模仿了人脑神经系统处理信息，存储以及检索的功能，它是一个非线性的映射系统。第6章和第7章介绍了两种数据挖掘中常用的传统统计的方法，回归分析和时间序列。回归分析是寻找自变量和因变量之间关系的预测模型，包括线性回归和Logistic回归；时间序列分析，顾名思义，是处理以时间为序的观测数据的方法。本书的一个特点是不仅对上述方法作了理论的阐述，还结合案例分析讲述了如何应用STATISTICA软件实现上述方法对数据的分析，是一本理论和实践相结合的理论性和应用性都很强的书。

在本书的编写过程中，中国人民大学统计学院的研究生参与了初稿的部分内容写作，再次对他们的辛勤工作表示衷心的感谢。他们是刘冬、戴杭君、张倩、孙兆楠、刘中华、詹瑾和王曦。此外要特别感谢中国人民大学统计学院的易丹辉教授，她对本书的写作提出了非常多的宝贵意见。

为便于读者学习，我们将本书练习题和案例部分的数据放在人大经管在线（www.rdjg.com.cn）上，读者可免费下载。

由于编者的水平和时间有限，错误之处在所难免，恳请读者批评指正。

编者

目　录

数据挖掘概述

1.1 数据挖掘定义

数据挖掘是一个多学科交叉研究领域，不仅大学里的专门研究人员在使用它，商业公司的专家和技术人员也在密切地关注它；它不仅涉及人工智能领域以及统计学的应用，而且也涉及数据库的使用。不同领域的人从不同的研究背景出发研究不同行业的数据，也就给了数据挖掘不同的内容和定义。这里我们就数据挖掘的技术定义以及商业定义展开讨论。

1.1.1 数据挖掘的技术定义

数据挖掘（data mining，DM）就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，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。与数据挖掘相近的同义词有数据融合、数据分析和决策支持等。

这个定义包括以下几层含义：

（1）数据源必须是真实的、大量的、含噪声的；

（2）发现的是用户感兴趣的知识；

（3）发现的知识要可接受、可理解、可运用；

（4）并不要求发现放之四海而皆准的知识，仅解决特定的问题。

这里原始数据可以是结构化的，如关系数据库中的数据；也可以是半结构化的，如文本、图形和图像数据；甚至可以是分布在网络上的异构型数据。知识则不但可以被用于信息管理、查询优化、决策支持和过程控制等，还可以用于数据自身的维护。更重要的是发现知识的方法，它可以是数学化的，也可以是非数学化的；可以是演绎的，也可以是归纳的。这些都充分体现了数据挖掘的多样性和广泛性，它把人们对数据的应用从低层次的简单查询提升到从数据中挖掘知识、提供决策的高端市场。

1.1.2 数据挖掘的商业定义

从应用的角度看，数据挖掘是一种新的商业信息处理技术，其主要特点是对商业数据库中的大量业务数据进行抽取、转换、分析和其他模型化处理，从中提取辅助商业决策的关键性数据。

现在，由于各行业业务自动化的实现，商业领域产生了大量的业务数据，这些数据不再是为了分析目的而收集，而是由于纯粹的商业运作而产生。分析这些数据也不再是单纯为了研究的需要，更主要的是为商业决策提供真正有价值的信息，进而获得利润。但所有企业面临的一个共同问题是：企业数据量非常大，而其中真正有价值的信息却很少，因此从大量的数据中经过深层分析，获得有利于商业运作、提高竞争力的信息也就显得愈加重要，数据挖掘就在此时大显神通。

数据挖掘可以描述为：按企业既定业务目标，对大量的企业数据进行探索和分析，揭示隐藏的、未知的或验证已知的规律，并进一步将其模型化的先进有效的方法。

数据挖掘在商业领域最成功的例子莫过于尿布与啤酒的关联。一家超市的老板经过分析其销售数据发现，一般年轻父亲在来超市给孩子买尿布的时候，总是喜欢捎带着买上两瓶啤酒。于是，超市老板把啤酒与尿布这两样风马牛不相及的商品摆放在一起，并且搭配着摆放了一些下酒的小菜，这样就极大地方便了顾客，同时也促进了啤酒的销量。如今，这家超市已经发展成美国第一大零售商，它就是沃尔玛公司。这个故事说明，在大量的数据中，总是隐藏着各种各样的信息，而数据挖掘正是用来从庞大的数据库中发掘这些有效信息的。所以有人形象地称“数据是知识和财富的源泉，从数据中挖掘知识就好像采矿一样”。

1.2　数据挖掘的重要性及意义

现代企业经常搜集了大量资料，包括市场、客户、供货商、竞争对手以及未来趋势等重要信息，但是信息超载与非结构化，使得企业决策者无法有效利用现存的信息，甚至使决策行为产生混乱与误用。

通过数据挖掘技术，从海量数据中挖掘出不同的信息与知识，作为决策支持之用，必能使企业产生竞争优势。数据挖掘不是为了替代传统的统计分析方法，相反，数据挖掘主要体现在利用统计分析技术和人工智能技术进行高级多元统计的研究和应用，是对这些方法的拓展和深化。

数据挖掘应用于企业，其重点在于企业领域方面的知识，而它的 Domain-specific Tools 要结合企业中使用者的专业知识和分析过程，才能发挥工具的效能并赢得竞争。换言之，就是要打破常规和超越平日的想象，展现企业目标与问题的知识精髓，以发现、提炼、解释一般人看不到、想不出的信息。

1. 企业应用数据挖掘是提升企业信息服务层次的需要

当今世界经济全球化的发展，使企业间的竞争日益激烈。企业如何以最短的时间、最快的速度、最少的投入赢得市场机遇，开发出用户接受的新产品，并以最快的方式销售产品，是企业在竞争中获胜的关键。要实现这一点，企业的各项行为应基于利用信息技术提供的快速、高效、智能化的信息服务，这给企业信息服务带来了新的挑战。具体来讲，包括信息服务的智能化、知识化、个性化和敏捷化。

信息服务智能化包括信息搜索智能化和决策分析智能化。信息搜索智能化是指信息搜索具有一定的知识推理和学习的能力，能根据用户需求和环境变化比较准确地理解用户意图，通过将复杂的、高难度的任务分析和分解，主动地、有针对性地向用户提供信息。同时能根据用户历史信息需求和用户行为，挖掘用户的兴趣爱好和意图，通过不断的训练学习，即使当用户需求没有明确给出或者不能准确表达时，也能推测用户的意图，主动搜集信息，向用户报告并提供信息。决策分析智能化是指利用决策支持和数据挖掘技术对集聚或集成的信息进行深层次分析处理，智能化地将信息转换成有用的决策知识，据此得到问题的总体特征，预测事物的发展趋势，针对某个决策问题向用户提供决策知识。

信息服务知识化是指企业信息服务人员向决策者提供的不是大量的信息，而是经过分析处理之后形成的有用的、集约化的知识。信息是事物的存在方式或运

动状态，以及对这种存在方式或运动状态的直接或间接描述；知识则是研究事物运动状态及其变化规律的结果，是经过加工处理系统化了的有用信息。因此，信息是关于“when，what，where，who”的描述，反映了事物的概况；而知识则是关于“why，how”的描述，它能够针对事物的因果关系进行预测，并指导用户开展下一步的工作。现代企业决策者最急切需求的不是大量信息，而是对信息加工分析之后形成的直接用来决策的知识。企业信息服务必须转变服务模式，从简单的信息处理转变为深层次的、知识化的信息分析。

信息服务的个性化表现在信息服务的精确性和主动性上。信息服务的精确性是指对每一个主体的特殊信息需求都能提供最对口的、专业性的、有针对性的信息。这要求企业信息系统能进行高效、专业、集成化的信息过滤，从而向信息用户提供个性化的信息服务。信息服务的主动性是指通过采用自适应方法，动态地从用户使用信息资源的记录中分析用户的个性、兴趣、心理和使用习惯等信息，获取用户的真正信息需求，主动向用户提供可能需要的信息。

信息服务的敏捷化是指对不同信息需求有快速的反应能力和灵敏的适应能力，无论环境怎么变化都能够及时满足用户的动态信息需求。对环境变化的被动适应能力和主动创新能力是信息系统具备敏捷性的两种重要能力。被动适应能力是指通过对信息系统内部进行重构使其能及时适应变化，在用户需要哪方面信息时就能及时、准确地向其提供。主动创新能力是指信息系统能正确预测未来的各种变化趋势，能根据历史数据分析和学习用户可能的信息需求，在用户尚未认识到其需求时，就可以向用户提供其内心想要的信息；还可以对用户表述出的信息需求提供前瞻性、预测性的分析信息。现代企业在激烈竞争中获胜的法宝就是对市场的快速反应能力和灵敏的适应能力。

要应对企业信息服务的挑战，需要建立以计算机技术、网络技术、数据仓库技术、联机分析处理技术和数据挖掘技术为基础的高层次信息分析处理技术。数据挖掘可以使企业信息服务从单纯的信息收集、存储、整理、利用、变无序信息为有序信息向信息创新、信息整合、信息再生产、变信息为知识的深层次加工方向发展。

2. 企业应用数据挖掘是企业信息化的需要

随着企业管理现代化和信息技术的发展，企业信息化已经成为企业适应竞争环境、发展壮大的有力武器。企业信息化也作为一项基本国策在全国范围内推行。目前已经有很多大型企业相继实施了企业信息化，如联想、海尔、斯达、神州数码等，并且在实践中取得了很好的效益；对于中小企业，虽然现在完全实施信息化的条件还不成熟，但它们大部分也都认识到了信息化对企业生存发展的重

要作用，通过企业信息基础设施的建设逐步为完全信息化提供良好的基础条件。

随着企业信息化观念和实践的推进，企业业务操作流程基本上都实现了自动化，很多企业的管理过程也实现了无纸化和智能化，企业信息化已经成为企业发展的重要动力和发展方向。目前很多企业已从简单的批处理、联机事务处理的信息处理时代迈入了联机分析处理、数据仓库和数据挖掘的信息分析时代。

3. 企业应用数据挖掘是企业信息共享的需要

现代任何一个企业在发展的过程中都非常注重团队精神，注重公司和谐统一的发展，希望企业能快速地应对内外部环境的变化，希望企业能同其供应商、零售商等合作伙伴实现“共赢”。企业的这些关注点和希望可以说是一种非常普遍的问题。其实在这些问题的背后有一个基本的保证，就是信息共享。如果企业没有信息共享的理念、机制和相应的信息技术支持，这些问题都不可能得到很好的解决。

但是随着信息技术在企业的应用及企业所处环境的日益复杂多变，企业数据的来源具有广泛性和分散性，各种不同来源的数据在格式上存在很大的差别，并产生大量的冗余，难以进行比较、鉴别和利用；同时这些数据来源和存储不规范，不具统一性，难以从这些不同的数据中得出统一的结论。在这些因素的影响下，企业虽然建立了内部网，但在企业内部却形成了以部门为单位的一个个数据和信息孤岛，难以实现真正的信息共享。如何共享分散在异构环境中的数据源，及时得到准确信息已经成为企业日常经营管理顺利进行的一个关键因素，这时就需要借助于新的数据分析处理技术——数据挖掘技术。

4. 企业应用数据挖掘是企业决策的需要

企业决策是企业管理中最重要的一部分，企业管理就是决策，决策失误是企业管理的最大失误。一次错误的决策将会对企业产生巨大的甚至是毁灭性的打击。决策从来都是企业非常关注并下大力气去解决的问题。但是随着市场竞争的日益激烈和复杂化，影响企业决策的因素越来越多，这些因素之间的关系也越来越复杂。企业要想在竞争中取胜，必须深层次地挖掘、分析企业积聚和采集到的大量当前和历史的生产经营数据、合作伙伴的数据、竞争对手的数据、客户数据以及相关环境的数据，自动快速地从中获取对决策有用的信息，为决策者提供快速、准确、方便的决策支持。随着经济一体化和全球化的发展，大规模、超大规模的公司越来越多，企业数据库日渐庞大，跨地域甚至跨国界的实时信息存储调用越发常见，企业对系统整合与知识更新的需求越来越强烈，决策所涉及的信息的分析处理远非人力或传统信息分析方法所能胜任，企业必须使用具有强大数据分析处理能力的技术如决策支持系统等进行无缝集成，及时处理这些海量的信

息，以利于决策者做出更加个性化和有针对性的决策。

1.3 数据挖掘功能

数据挖掘的功能一般可以分为两大类：描述和预测。描述类挖掘任务刻画了数据库中数据的一般特性；预测类挖掘任务在当前数据上进行分析，以此进行推断。

一般而言，数据挖掘的功能与挖掘的目标数据类型是相关的。某些功能只能应用在某种特定的数据类型上，而某些功能则可以应用在多种不同类型的数据库上。了解数据挖掘的分类，理解被挖掘的对象，并在此基础上对挖掘对象按挖掘功能进行分类，有助于我们按照用户需求选择合适的挖掘算法或挖掘工具来辅助企业制定决策，同时也是我们准确地分析问题和解决问题的依据。

1. 概念描述

概念描述（concept description）就是通过对某类对象关联数据的汇总、分析和比较，对此类对象的内涵进行描述，并概括这类对象的有关特征。数据挖掘中的概念描述主要关心从数据泛化的角度来讨论数据总结，或者说就是把数据库中的有关数据从低层次抽象到高层次上的过程。

概念描述分为特征性描述和区别性描述，前者描述某类对象的共同特征，生成一个类的特征性描述，只涉及该类对象中所有个体的共性；一般采用饼图、柱状图、曲线、多维数据立方体、含交叉表的多维表来输出描述结果，还可以采用概化关系或规则形式表示。后者描述不同类别对象之间的区别，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，而这种比较必须在具备可比性的两个或多个类之间进行。

2. 多层次概念描述

一般的，由数据归纳出的概念是有层次的，如 location 是“中国人民大学”，那么我们可能通过背景知识（background knowledge）归纳出“北京市”、“中国”、“亚洲”等不同层次的更高级的概念，这就是多层次概念描述——概念描述中的探索性问题。

所谓概念分层其实就是将低层概念集映射到高层概念集的方法。例如，一个记录销售人员销售情况的数据库的表（SALES NO.，NAME，AGE，VALUE，DEPT），它的每个属性的定义域都可能存在蕴含于领域知识内的概念延伸，如所在部门 DEPT 可能在特定的条件下需要知道所在公司（COMPANY）城市（CITY）

或国家（COUNTRY），因为在更高层次的数据综合和分析是决策的基础。

3. 关联分析

数据关联（association analysis）是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。关联可分为简单关联、时序关联、因果关联。关联分析的目的是找出数据库中隐藏的关联网，但是很多时候我们并不知道数据库中数据的关联函数，即使知道也是不确定的，因此关联分析生成的规则带有可信度。

关联规则的一个很浅显的表述是形如“面包、黄油→牛奶”［可信度 90%］的一种规则，它包含的意思是“在购买面包和黄油的顾客当中，有 90%的人同时也买了牛奶”。显然，这种关联规则反映了顾客的购买习惯。如果商家能够充分利用这些购买习惯，就可以增加商品的销量，提高销售利润。

目前，关联规则主要是针对事务性数据库，特别是售货数据，由于条形码技术的发展，零售部门可以利用前端收款机收集存储大量的售货数据，如果对这些历史事务数据进行分析，则可对顾客的购买行为提供极有价值的信息。例如，可以帮助摆放货架上的商品，规划市场、减小库存，对市场变化提供预测，等等。

4. 聚类

将物理或抽象对象的集合分组成为由类似的对象组成的多个类的方法就是聚类（clustering），它没有给定分类的任何依据（如预定的分类表、预订的类目），是一种根据信息相似度进行数据聚集的方法。聚类的目的是根据一定的规则，对数据进行合理的分组，并用显式或隐式的方法描述不同的类别。由于分析可以采用不同的算法，所以对于相同的数据集合可能有不同的划分。

5. 分类

分类（classification）在数据挖掘中是一项非常重要的任务，其目的是找出一组能够描述数据集合典型特征的模型或者函数，以便能够识别位置数据的归属或类别。实质上，数据分类就是从数据库对象中发现共性，并将数据对象分成不同类别的一个过程。首先要对训练数据进行分析，使用数据的某些特征属性，给出每个类的准确描述，即分类规则，然后使用这些描述，对数据库中的其他数据进行分类。

简单地说，分类过程包含两步：第一步，建立一个模型，描述指定的数据类集；第二步，使用模型进行分类。模型可以用多种形式来表示，如分类规则、判定树、数学公式或神经网络等。例如，通过训练数据得到规则：IF 年龄＝

“31…40”AND 收入＝“较高”THEN 信用程度＝“优秀”，这个规则的含义就是年龄在 31～40 岁之间，收入较高的情况下，这类顾客群的信用程度被认为是“优秀”。

任何一种分类技术与算法都不是万能的，不同的商业问题，需要用不同的方法解决，即使对于同一个商业问题，也可能有多种分类算法。分类的效果一般和数据的特点相关，有些数据噪声大、有缺失值、分布稀疏，有些属性是离散的而有些是连续的，因此对于一个特定问题和一类特定数据，需要评估具体算法的适用性。

6. 偏差检测

数据库中的数据常有一些异常记录，从数据库中检测这些偏差很有意义。偏差包括很多潜在的知识，如分类中的反常实例、不满足规则的特例、观测结果与模型预测值的偏差、量值随时间的变化等。偏差检测的基本方法是寻找观测结果与参照值之间有意义的差别。

7. 孤立点分析

数据库中可能包括一些数据对象，它们与数据的一般行为或模型不一致，这些数据被称作孤立点（outlier），大部分的数据挖掘方法将孤立点视为噪声或异常而丢失。然而，在一些应用中（如欺骗检验），罕见的时间可能比正常出现的时间更有趣，这就需要孤立点分析。孤立点分析可以使用统计实验检测，假定一个数据分布或概率模型，并使用距离度量，到所有聚类的距离均很大的对象被视为孤立点。此外，基于偏差检测的方法是通过考察若干对象主要特征的差别识别孤立点，而不是使用统计或距离度量。孤立点分析可以用于发现信用卡欺骗事例，通过检测一个给定账号，与正常付费相比，以付款数额特别大来发现信用卡欺骗性使用。

8. 自动预测趋势和行为

主要是针对那些具有时序（time series）属性的数据，如股票价格等，或者是序列项目（sequence items）的数据，如年龄和薪水对照等，发现长期的趋势变化。有许多来自统计学的方法经过改造可用于数据挖掘，如基于 n 阶移动平均值数据挖掘自动在大型数据库中寻找预测性信息，以往需要进行大量手工分析的问题如今可以迅速直接由数据本身得出结论。一个典型的例子是市场预测问题，数据挖掘使用过去有关促销的数据来寻找未来投资中回报最大的用户，其他可预测的问题包括预报破产以及认定对指定事件最可能做出反应的群体。

9. 时序演变分析

数据的时序演变分析（temporal evolution analysis）是寻找事件或对象行为