普通高等教育“十三五”规划教材

PUTONG GAODENG JIAOYU "13·5" GUIHUA JIAOCAI

数据挖掘学习方法

王 玲 编著

冶金工业出版社
www.cnmip.com.cn

普通[redacted]”规划教材

数据挖掘学习方法

王 玲 编著

北 京
冶 金 工 业 出 版 社
2017

内 容 提 要

本书系统地介绍了数据挖掘的方法和技术，主要内容包括：决策树挖掘；关联规则挖掘；逻辑回归；神经网络；聚类分析；支持向量机；降维；异常检测等。每一章都会涉及学习要点、学习难点和思考题，希望能使学生对数据挖掘的整体结构、理论、概念、技术和方法有深入的认识和了解；掌握数据挖掘的技术、方法及数据挖掘应用系统开发，了解数据仓库和数据挖掘技术的研究问题、现状及未来的研究方向。并且结合具体案例的分析，实现数据挖掘的功能。希望学生在创新意识、科研能力等方面得到提高。

本教材可供自动化及相关专业本科生及研究生使用，也可供从事自动化技术的科技人员参考。

图书在版编目(CIP)数据

数据挖掘学习方法/王玲编著. —北京：冶金工业出版社，2017.8

普通高等教育“十三五”规划教材

ISBN 978-7-5024-7545-1

Ⅰ.①数… Ⅱ.①王… Ⅲ.①数据采集—高等学校—教材 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字（2017）第 157867 号

出 版 人 谭学余

地 址 北京市东城区嵩祝院北巷 39 号 邮编 100009 电话 (010)64027926

网 址 www.cnmip.com.cn 电子信箱 yjcbs@cnmip.com.cn

责任编辑 戈 兰 夏小雪 美术编辑 彭子赫 版式设计 孙跃红

责任校对 石 静 责任印制 李玉山

ISBN 978-7-5024-7545-1

冶金工业出版社出版发行；各地新华书店经销；北京印刷一厂印刷

2017 年 8 月第 1 版，2017 年 8 月第 1 次印刷

787mm×1092mm 1/16；9.25 印张；218 千字；135 页

32.00 元

冶金工业出版社 投稿电话 (010)64027932 投稿信箱 tougao@cnmip.com.cn

冶金工业出版社营销中心 电话 (010)64044283 传真 (010)64027893

冶金书店 地址 北京市东四西大街 46 号(100010) 电话 (010)65289081(兼传真)

冶金工业出版社天猫旗舰店 yjgycbs.tmall.com

前　言

数据挖掘是一门与数据库、统计学、机器学习等多学科交叉的新兴科学，旨在从数据中抽取隐含的、未知的和潜在有用的信息，在商业、金融、医学、科学研究、工程与政府部门管理都有广泛应用。一本详细设计、强调概念、技术丰富而平衡的数据挖掘教材将会为今后从事数据挖掘研究的学生研究、开发与使用数据挖掘技术提供好的指导方向。

根据数据挖掘课程的定位和教学基本要求，我们遵循实用、简单、够用的原则，控制教材篇幅，使本书具有良好的教学适用性，内容上注意知识的模块化与层次，给学生一个基本知识范畴，相对于课时的增减使之具有较大弹性。让学生在掌握基础知识的同时，也能够了解它的具体应用。

本书梳理了数据挖掘的多种研究方法，注重领域核心方法的论述，知识点比较广泛，叙述简明、语言准确，有助于增强学生的个性化学习与自学能力，调动学生的学习主动性而不会让学生不堪重负或望而生畏。全书共14章，第1章介绍数据挖掘的概述；第2章介绍了数据仓库原理以及数据仓库设计过程；第3章~第13章分别从聚类、关联规则、决策树、逻辑回归、多变量线性回归、神经网络、支持向量机、异常检测、推荐系统、大规模数据挖掘算法等多个主题讲述了算法和概念。第14章给出了一个具体的应用案例。这样做的目的是使教材能够比较全面地覆盖数据挖掘方法的基本知识点，言简意赅地提炼隐藏在其中的数学思想方法，以求给学生以启发和引导，有助于贯彻创新教育教学理念。

参加本书编写工作的还有孟建瑶、徐培培、郭辉等。由于作者的水平有限，加之编写时间仓促，书中不妥之处，恳请读者批评指正。

王　玲

2017年5月于北京

冶金工业出版社部分图书推荐

书　名	作　者	定价(元)
微机原理及接口技术习题与实验指导（高等教材）	董　洁　等主编	46.00
工业自动化生产线实训教程（高等教材）	李　擎　等主编	38.00
过程控制（高等教材）	彭开香　主编	49.00
自动检测技术（第3版）（高等教材）	李希胜　等主编	45.00
流体仿真与应用（高等教材）	刘国勇　编著	49.00
职业卫生工程（高等教材）	杜翠凤　等编著	38.00
热轧生产自动化技术（第2版）	刘　玠　等编著	118.00
冷轧生产自动化技术（第2版）	孙一康　等编著	78.00
冶金企业管理信息化技术（第2版）	许海洪　等编著	68.00
炉外精炼及连铸自动化技术（第2版）	蒋慎言　编著	96.00
炼钢生产自动化技术（第2版）	蒋慎言　等编著	108.00
智能节电技术	周梦公　编著	96.00
刘玠文集	文集编辑小组　编	290.00
钢铁生产控制及管理系统	骆德欢　等主编	88.00
钢铁企业电力设计手册（上册）	本书编委会　编	185.00
钢铁企业电力设计手册（下册）	本书编委会　编	190.00
变频器基础及应用（第2版）	原　魁　等编著	29.00
钢铁工业绿色工艺技术	于　勇　等编著	146.00
铁矿石优化配矿实用技术	许满兴　等编著	76.00
稀土永磁材料（上、下册）	胡伯平　等编著	260.00
稀土在低合金及合金钢中的应用	王龙妹　著	128.00
煤气安全作业应知应会300问	张天启　主编	46.00
钢铁材料力学与工艺性能标准试样图集及加工工艺汇编	王克杰　等主编	148.00
安全技能应知应会500问	张天启　主编	38.00
走进黄金世界	胡宪铭　等编著	76.00
钢铁企业风险与风险管理	牟宝喜　主编	106.00

目　录

第1章　数据挖掘概述

教学要求： 了解数据挖掘技术，掌握数据挖掘的概念；
了解数据仓库的发展及展望，掌握数据仓库的概念；
掌握数据挖掘和在线分析处理的关系；
了解数据挖掘与机器学习、统计之间的关系；
掌握数据挖掘的研究内容。

重　　点： 数据挖掘的概念；
数据仓库的概念；
数据挖掘研究内容。

难　　点： 数据挖掘与相关技术之间的关系；
数据挖掘的研究内容。

第1章　课件

随着数据库技术的迅速发展以及数据库管理系统的广泛应用，人们积累的数据越来越多。激增的数据背后隐藏着许多重要的信息，人们希望能够对其进行更高层次的分析，以便更好地利用这些数据。目前的数据库系统可以高效地实现数据的录入、查询、统计等功能，但无法发现数据中存在的关系和规则，无法根据现有的数据预测未来的发展趋势，缺乏挖掘数据背后隐藏的知识手段，导致了“数据爆炸但知识贫乏”的现象。

数据挖掘技术是人们长期对数据库技术进行研究和开发的结果。起初各种数据是存储在计算机的数据库中的，然后发展到可对数据库进行查询和访问，进而发展到对数据库的即时遍历。数据挖掘使数据库技术进入了一个更高级的阶段，它不仅能对过去的数据进行查询和遍历，并且能够找出过去数据之间的潜在联系，从而促进信息的传递。

数据挖掘其实也是一个逐渐演变的过程，电子数据处理的初期，人们就试图通过某些方法来实现自动决策支持，当时机器学习成为人们关心的焦点。机器学习的过程就是将一些已知的并已被成功解决的问题作为范例输入计算机，机器通过学习这些范例总结并生成相应的规则，这些规则具有通用性，使用它们可以解决某一类的问题。随后，随着神经网络技术的形成和发展，人们的注意力转向知识工程，知识工程不同于机器学习那样给计算机输入范例，让它生成出规则，而是直接给计算机输入已被代码化的规则，而计算机是通过使用这些规则来解决某些问题。专家系统就是这种方法所得到的成果，但它有投资大、效果不甚理想等缺点。20世纪80年代人们又在新的神经网络理论的指导下，重新回到机器学习的方法上，并将其成果应用于处理大型商业数据库。随着在80年代末一个新的术语，它就是数据库中的知识发现，简称KDD（Knowledge Discovery in Database）。它泛指所有从源数据中发掘模式或联系的方法，人们接受了这个术语，并用KDD来描述整个数据

发掘的过程，包括最开始的制定业务目标到最终的结果分析，而用数据挖掘（Data Mining）来描述使用挖掘算法进行数据挖掘的子过程。但最近人们却逐渐开始使用数据挖掘中有许多工作可以由统计方法来完成，并认为最好的策略是将统计方法与数据挖掘有机地结合起来。

数据挖掘的核心模块技术历经了数十年的发展，其中包括数理统计、人工智能、机器学习。今天，这些成熟的技术，加上高性能的关系数据库引擎以及广泛的数据集成，让数据挖掘技术在当前的数据仓库环境中进入到了实用的阶段。

1.1 数据挖掘的定义及含义

数据挖掘（Data Mining）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。这个定义包括好几层含义：数据源必须是真实的、大量的、含噪声的；发现的是用户感兴趣的知识；发现的知识要可接受、可理解、可运用；并不要求发现放之四海皆准的知识，仅支持特定的发现问题。

何为知识？从广义上理解，数据、信息也是知识的表现形式，但是人们更把概念、规则、模式、规律和约束等看作知识。人们把数据看做是形成知识的源泉，好像从矿石中采矿或淘金一样。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；甚至是分布在网络上的异构型数据。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。发现的知识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。因此，数据挖掘是一门交叉学科，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，提供决策支持。在这种需求牵引下，汇聚了不同领域的研究者，尤其是数据库技术、人工智能技术、数理统计、可视化技术、并行计算等方面的学者和工程技术人员，投身到数据挖掘这一新兴的研究领域，形成新的技术热点。这里所说的知识发现，不是要求发现放之四海而皆准的真理，也不是要去发现崭新的自然科学定理和纯数学公式，更不是什么机器定理证明。实际上，所有发现的知识都是相对的，是有特定前提和约束条件，面向特定领域的，同时还要能够易于被用户理解。最好能用自然语言表达所发现的结果。

1.2 数据挖掘的作用

这个显然跟你的业务有关系，你的业务需要不需要，哪里需要，需要干什么，希望投入什么样的成本，产出什么样的结果，这是决定你要不要做以及怎么做的一个基本考虑。大数据的核心是对数据的应用，之所以用大数据，就是希望通过数据分析处理，来更精准地把握用户、客户行为和更好地挖掘信息的价值，提升业务的利润和控制成本。

（1）大数据挖掘可以让杂乱无序的数据清晰化、可用度高。大数据有两个典型特征，其一是数据量大，其二是计算复杂。与传统数据库相比，大数据的结构化程度、可用度、

数据抽取、数据清洗都是很大的一块工作。

特别典型的传统生产销售型企业的业务系统数据是隔离、分裂的，有销售的、生产的、财务的、客户的等，不同方面其实都是为自己负责的业务目标和输出构建自己的IT系统、甚至是外包给不同的IT集成商或者软件开发商做的，因而系统都是相对独立，这种独立的结果不只是隔离，而是从数据的结构、数据的记录与存储、软件系统负载等产品技术层面都不尽同。数据挖掘需要根据你的目标构建挖掘模型，建立起多个数据系统的关联。

（2）让数据和数据之间发生关系，这关系可能产生化学反应。著名的啤酒与尿布、口香糖与避孕套的例子就是典型的数据之间隐性关系的发现，通过对消费行为数据进行建模和分析，能够发现两个原本不相干的东西，在用户采购东西的时候发生了关系，那么针对这一发现优化你的货架物品摆放就能够提高销售量。用过亚马逊的朋友可能都看到过，买个手机马上推荐跟手机壳、存储卡打包购买有折扣哦，这种推荐能节省用户的成本。

（3）对数据产生状态进行监控，发现异常，预警纠错。通过对系统产生的数据按照时间建模，记录每个时间点、时间周期内的均值和上下区间，如果某个节点出现超乎寻常的状况，系统能很快发现问题并进行预警和排查。当然这只是技术系统的价值。

从业务系统上，这种数据异常将会给你的经营状况给出警示，帮你从历史时间维度对比，判断事情变化的因由，提供你决策分析必要的时间、数据和关联信息参考。

（4）通过数据挖掘建立知识模型，提供决策支持信息。信息系统发挥更大的价值在于能通过信息的整合，帮你提供决策参考信息。以前有一个提法叫做知识发现KDD，随着互联网信息内容的丰富、UGC分众智慧的发挥，网络信息的价值效用也越来越大。通过信息存在和信息特征提取，建立起不同信息之间的关联，并能通过语义分析、情感分析，提炼出信息本身的价值倾向、态度、消费效用等，这将为信息在决策参考上提供更系统、数据化的分析和参考。

（5）强大的数据处理和分析能够建立以数据驱动的垂直商业生态。数据挖掘的技术系统将负责将所有数据，按照目标重新梳理和建立跟模型对应的数据索引。这个重新构建数据的秩序将大大增加数据的可用性。从垂直行业切入，针对这行业信息服务的需求，建立模型，并不断优化各个细节和子节点的输出，使得行业参与的各角色能在生态上获取自己的利益和价值，那么这将建立起针对这个细分行业的垂直业务生态。我们身边已经有很多大规模数据的应用，比如电商购物对用户做推荐，基于用户群和用户行为的分类做精准的广告投放等，亦或计算气象预报，计算地质数据做石油探测、矿产探测，还有金融行业对投资、贷款等的风险预估。跟大规模数据挖掘相关的主要技术有数据存储、数据挖掘的分布式计算平台，结构化存储，计算任务管理和调度等，所以一般性的大数据挖掘项目都跟云计算、云存储和自动运维系统密切相关，需要一定投入才能搞得定。

1.3 数据挖掘和数据仓库

大部分情况下，数据挖掘都要先把数据从数据仓库中拿到数据挖掘库或数据集市中（如图1-1所示）。从数据仓库中直接得到进行数据挖掘的数据有许多好处。就如我们后面

会讲到的，数据仓库的数据清理和数据挖掘的数据清理差不多，如果数据在导入数据仓库时已经清理过，那很可能在做数据挖掘时就没必要再清理一次了，而且所有的数据不一致的问题都已经被你解决了。从数据仓库中提取的数据挖掘集市、数据挖掘库可能是你的数据仓库的一个逻辑上的子集，而不一定非得是物理上单独的数据库。但如果你的数据仓库的计算资源已经很紧张，那你最好还是建立一个单独的数据挖掘库。

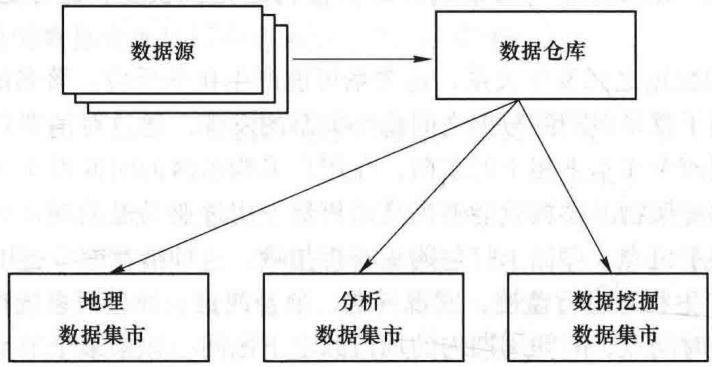

图 1-1 从数据仓库中提取的数据挖掘集市

当然，为了进行数据挖掘你也不必非得建立一个数据仓库，数据仓库不是必需的。建立一个巨大的数据仓库，把各个不同源的数据统一在一起，解决所有的数据冲突问题，然后把所有的数据导到一个数据仓库内，是一项巨大的工程，可能要用几年的时间花上百万的钱才能完成。只是为了数据挖掘，你可以把一个或几个事务数据库导到一个只读的数据库中，就把它当做数据集市（如图 1-2 所示），然后在它上面进行数据挖掘。

图 1-2 从操作数据库中提取的数据挖掘集市

1.4 数据挖掘和在线分析处理

一个经常问的问题是，数据挖掘和在线分析处理（OLAP）到底有何不同。下面将会解释，他们是完全不同的工具，基于的技术也大相径庭。

OLAP 是决策支持领域的一部分。传统的查询和报表工具是告诉你数据库中都有什么（What happened），OLAP 则更进一步告诉你下一步会怎么样（What next）和如果我采取这样的措施又会怎么样（What if）。用户首先建立一个假设，然后用 OLAP 检索数据库来验证这个假设是否正确。比如，一个分析师想找到什么原因导致了贷款拖欠，他可能先做一个初始的假定，认为低收入的人信用度也低，然后用 OLAP 来验证他这个假设。如果这个假设没有被证实，他可能去察看那些高负债的账户，如果还不行，他也许要把收入和负债一起考虑，一直进行下去，直到找到他想要的结果或放弃。

也就是说，OLAP 分析师是建立一系列的假设，然后通过 OLAP 来证实或推翻这些假设来最终得到自己的结论。OLAP 分析过程在本质上是一个演绎推理的过程。但是如果分析的变量达到几十或上百个，那么再用 OLAP 手动分析验证这些假设将是一件非常困难和痛苦的事情。

数据挖掘与 OLAP 不同的地方是，数据挖掘不是用于验证某个假定的模式（模型）的正确性，而是在数据库中自己寻找模型。他在本质上是一个归纳的过程。比如，一个用数据挖掘工具的分析师想找到引起贷款拖欠的风险因素。数据挖掘工具可能帮他找到高负债和低收入是引起这个问题的因素，甚至还可能发现一些分析师从来没有想过或试过的其他因素，比如年龄。

数据挖掘和 OLAP 具有一定的互补性。在利用数据挖掘出来的结论采取行动之前，你也许要验证一下如果采取这样的行动会给公司带来什么样的影响，那么 OLAP 工具能回答你的这些问题。而且在知识发现的早期阶段，OLAP 工具还有其他一些用途。可以帮你探索数据，找到哪些是对一个问题比较重要的变量，发现异常数据和互相影响的变量。这都能帮你更好的理解你的数据，加快知识发现的过程。

1.5 数据挖掘、机器学习和统计

数据挖掘利用了人工智能（AI）和统计分析的进步所带来的好处。这两门学科都致力于模式发现和预测。

数据挖掘不是为了替代传统的统计分析技术。相反，他是统计分析方法学的延伸和扩展。大多数的统计分析技术都基于完善的数学理论和高超的技巧，预测的准确度还是令人满意的，但对使用者的要求很高。而随着计算机计算能力的不断增强，我们有可能利用计算机强大的计算能力只通过相对简单和固定的方法完成同样的功能。

一些新兴的技术同样在知识发现领域取得了很好的效果，如神经元网络和决策树，在足够多的数据和计算能力下，他们几乎不用人的关照自动就能完成许多有价值的功能。数据挖掘就是利用了统计和人工智能技术的应用程序，它把这些高深复杂的技术封装起来，使人们不用自己掌握这些技术也能完成同样的功能，并且更专注于自己所要解决的问题。

1.6 软硬件发展对数据挖掘的影响

使数据挖掘这件事情成为可能的关键一点是计算机性能价格比的巨大进步。在过去的几年里磁盘存储器的价格几乎降低了 99%，这在很大程度上改变了企业界对数据收集和存储的态度。如果每兆的价格是 10 元，那存放 1TB 的价格是 10000000 元，但当每兆的价格降为 0.1 元时，存储同样的数据只有 100000 元。计算机计算能力价格的降低同样非常显著。每一代芯片的诞生都会把 CPU 的计算能力提高一大步。内存 RAM 也同样降价迅速，几年之内每兆内存的价格由几百块钱降到现在只要几块钱。通常 PC 都有 64M 内存，工作站达到了 256M，拥有上 G 内存的服务器已经不是什么新鲜事了。

在单个 CPU 计算能力大幅提升的同时，基于多个 CPU 的并行系统也取得了很大的进步。目前几乎所有的服务器都支持多个 CPU，这些 SMP 服务器簇甚至能让成百上千个

CPU 同时工作。

基于并行系统的数据库管理系统也给数据挖掘技术的应用带来了便利。如果你有一个庞大而复杂的数据挖掘问题要求通过访问数据库取得数据，那么效率最高的办法就是利用一个本地的并行数据库。

所有这些都为数据挖掘的实施扫清了道路，随着时间的延续，我们相信这条道路会越来越平坦。

1.7 数据挖掘的类型和研究内容

随着数据挖掘研究逐步走向深入，它已经形成了三根强大的技术支柱：数据库、人工智能和数理统计。现代的数据挖掘主要包括描述性数据挖掘和预测性数据挖掘。描述性数据挖掘是以简洁概要的方式描述数据，并提供数据的有趣的一般性质。预测性数据挖掘是通过分析建立一个或者一组模型，并试图预测新数据集合的行为。下面分别介绍描述性数据挖掘和预测性数据挖掘的研究内容。

1.7.1 描述性数据挖掘

1.7.1.1 广义知识

在建立模型之前，首先要了解数据，获得广义知识，即类别特征的概括性描述知识。根据数据的微观特性发现其表征的、带有普遍性的、较高层次概念的、中观和宏观的知识，反映同类事物共同性质，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实现某些常用的代价较高的聚集函数的计算，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 SimonFraser 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。

1.7.1.2 聚类

聚类的目的是把数据对象分成各个聚类，各个簇，但聚类与分类也有显著的不同，聚类分析是一种无指导的学习，而分类的训练样本集的类标号是已知的，通过学习对训练数据集得出一个分类规则，再利用分类规则判定某个未知数据的类标号，分类是有指导的学习。进行聚类时，不存在类标号已知的训练数据集，没有什么模型可参考，聚类算法必须自己总结出各个聚类或簇之间的区别，根据某种规则来对数据对象进行聚类或分类，这个角度上讲，聚类是无指导的学习，它的算法本身远比分类的复杂度要高。目前，聚类分析有很多相应的算法，它其实是一个多学科的融合，大部分的聚类算法都是基于距离的聚类，这个距离是依据统计学中相关的公式和知识。对于聚类分析，各个其他领域也有着比

较深入的研究，像生物、医学等都对聚类分析有相应的研究和贡献。

1.7.1.3 关联分析

关联分析是一种探索数据的描述性方法，这些数据可以帮助识别数据库中数值之间的关系。它反映一个事件和其他事件之间依赖或关联的知识。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。关联规则的发现可分为两步。第一步是迭代识别所有的频繁项目集，要求频繁项目集的支持率不低于用户设定的最低值；第二步是从频繁项目集中构造可信度不低于用户设定的最低值的规则。识别或发现所有频繁项目集是关联规则发现算法的核心，也是计算量最大的部分。

1.7.2 预测性数据挖掘

预测性数据挖掘的目的是通过分析建立一个或一组模型，并试图预测新数据的行为。在从多种来源搜集数据的基础上，它通过构建现实世界的模型来实现，这些来源可包括企业交易，顾客历史和人口统计信息，过程控制数据，以及相关的外部数据库，例如银行交易信息或气象数据。模型建立的结果是对那些能用来进行有效预测的数据中的模式和关系的描述。

确定了预测目标之后，下一步是决定最合适的预测类型：（1）分类：预测行为属于什么类别或等级，或（2）回归：预测变量会有什么数值（如果它是随时间变化的变量，这就是所谓的时间序列预测）。现在，你可以选择模型类型：用神经网络来进行回归分析，以及可能用决策树来进行分类。也有传统的统计模型可供选择，如逻辑回归，判别分析，或一般线性模型。数据挖掘中最重要的模型类型将在后续章节中进行描述。

在预测模型中，我们的预测值或类被称为响应，相关或目标变量。用于建立或者训练预测模型使用的数据是已知变量响应的数值。这种训练有时被称为监督学习，因为被计算或估计值会与已知的结果进行比较（相反，在上一节中的描述性技术，如聚类，有时被称为无监督学习，因为没有已知的结果来引导算法）。

1.7.2.1 分类

分类是预测分类标号。什么是分类标号呢？我们知道属性值有两种基本的属性值，一种是分类属性，一种是量化属性。分类属性也叫离散属性，它的值是分成固定的区间之内的，是离散的值，而量化属性对应的是连续的值，根据分类时所对应的是离散的属性还是量化的属性，就可以把分类挖掘分成分类和预测两种类型。分类预测的是分类编号，根据训练数据集和类标号属性构建模型来分类新数据，这里主要包括两个过程，一个是构建模型来分类现有的数据，第二个是利用已有的模型对新数据进行分类。

最为典型的分类方法是基于决策树的分类方法。它是从实例集中构造决策树，是一种有指导的学习方法。该方法先根据训练子集（又称为窗口）形成决策树。如果该树不能对所有对象给出正确的分类，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。最终结果是一棵树，其叶结点是类名，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。

数据分类还有逻辑回归、线性判别分析、神经网络、粗糙集（RoughSet）等方法。