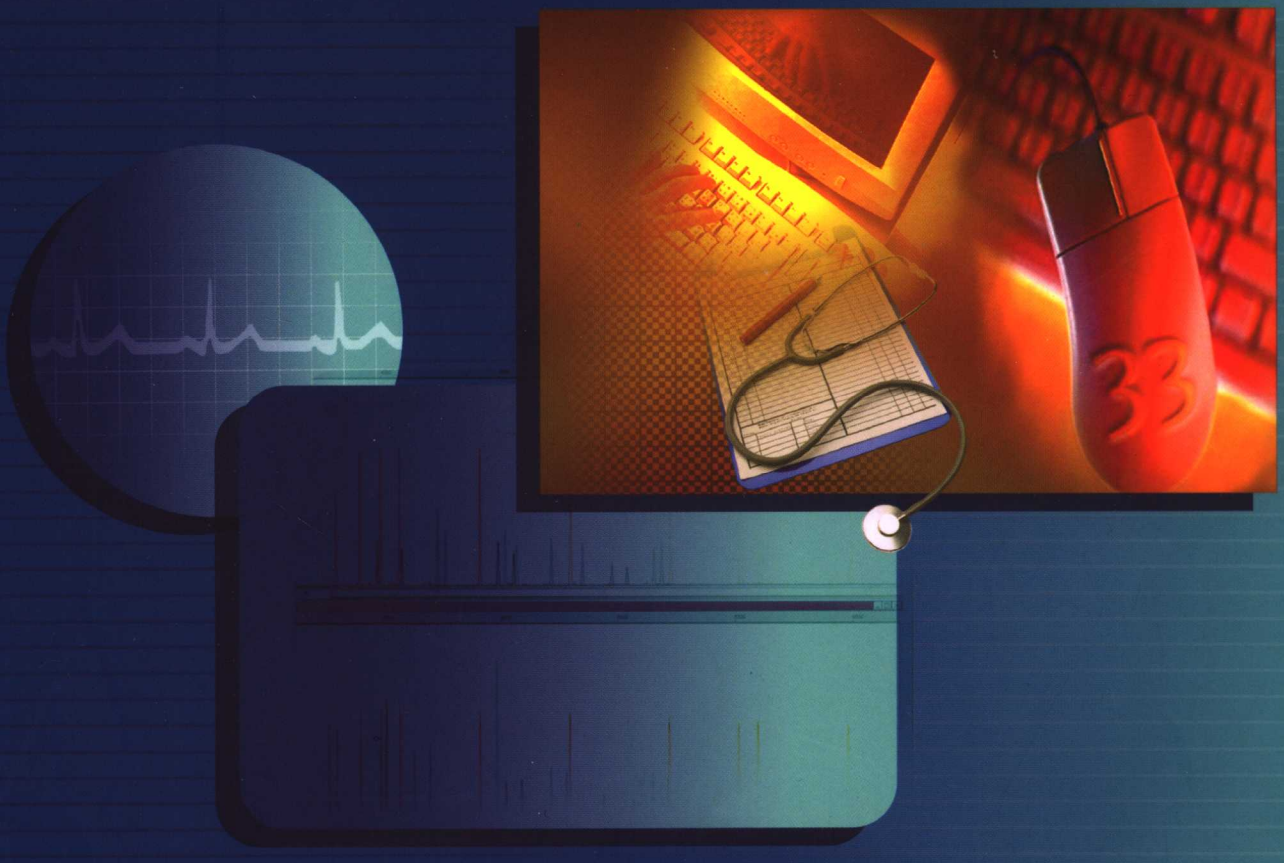
全国高等学校医学规划教材
（供信息管理与信息系统专业用）

医学数据挖掘

主编　崔　雷



高等教育出版社
Higher Education Press

全国高等学校医学规划教材

（供信息管理与信息系统专业用）

医学数据挖掘

主　编　崔　雷

副主编　刘建炜　马敬东　赵文龙

编　者（以姓氏拼音为序）

崔　雷　中国医科大学

胡德华　中南大学

刘建炜　中南大学

马敬东　华中科技大学

闫　雷　中国医科大学

叶明全　皖南医学院

袁永旭　山西医科大学

张　晗　中国医科大学

赵文龙　重庆医科大学

周宇葵　中南大学

高等教育出版社

Higher Education Press

内容简介

本书是国内第一部关于医学数据挖掘的教材。包括基础篇、核心篇和应用篇三个部分。基础篇介绍数据挖掘的基本概念和理论，核心篇介绍数据挖掘的主要算法和工具，应用篇则分别介绍数据挖掘在医学临床、分子生物学、预防医学、医院管理、文本和 Web 挖掘中的具体应用。

本书首先强调数据挖掘的基本概念和基本方法，重点介绍该领域的基本概念、基本过程和方法；各种算法以介绍其适用条件和原理为主，尽量少涉及具体算法的数学公式。其次，本书以应用为主，介绍数据挖掘方法在医学研究和服务中的应用实例，为学生今后进一步从事这一方面的深入研究提供基础。最后，本书在内容组织上力求全面系统，突出重点。由浅入深、突出交叉学科的特色的同时，注重所介绍知识的层次，适合不同水平读者的学习需要。

图书在版编目(CIP)数据

医学数据挖掘/崔雷主编.—北京：高等教育出版社，2006.7

ISBN 7-04-019078-8

Ⅰ.医… Ⅱ.崔… Ⅲ.数据采集-计算机应用-医学-教材 Ⅳ.R-05

中国版本图书馆 CIP 数据核字(2006)第 059269 号

策划编辑 刘晋秦　**责任编辑** 杨利平　**封面设计** 张　楠　**责任绘图** 朱　静
版式设计 马静如　**责任校对** 朱惠芳　**责任印制** 毛斯璐

出版发行	高等教育出版社	**购书热线**	010-58581118
社　　址	北京市西城区德外大街 4 号	**免费咨询**	800-810-0598
邮政编码	100011	**网　　址**	http://www.hep.edu.cn
总　　机	010-58581000		http://www.hep.com.cn
		网上订购	http://www.landraco.com
经　　销	蓝色畅想图书发行有限公司		http://www.landraco.com.cn
印　　刷	北京北苑印刷有限责任公司	**畅想教育**	http://www.widedu.com
开　　本	850×1168　1/16	**版　　次**	2006 年 7 月第 1 版
印　　张	14.5	**印　　次**	2006 年 7 月第 1 次印刷
字　　数	420 000	**定　　价**	23.40 元

物料号　19078-00

前　言

数据挖掘与知识发现是当前研究的热门领域，是集信息科学、管理科学、统计学和人工智能等学科于一身的交叉学科。数据挖掘在生物医学领域中的应用有着广阔的前景。对于医学信息管理专业而言，数据挖掘研究为信息管理通向知识管理架设了一座桥梁。

正是基于这个原因，许多医学院校的相关专业，如信息管理与信息系统、生物技术专业、计算机专业纷纷准备开设这门课程。但是，目前没有适合医学信息专业本科生水平的教材。为此，我们几家医学信息学专业的教师在高等教育出版社的组织下编写了这部教材。

本书包括基础篇、核心篇和应用篇三个部分。基础篇介绍数据挖掘的基本概念和理论，核心篇介绍数据挖掘的主要算法和工具，应用篇则分别介绍数据挖掘在医学临床、分子生物学、预防医学、医院管理、文本和 Web 挖掘中的具体应用。

本书的特点之一是强调基本概念和基本方法。由于数据挖掘和知识发现是目前新近兴起的一个研究领域，很多基本概念和基本理论尚处于发展完善之中，基于该研究目前在我国的发展状况以及目前医学信息管理专业本科生的知识结构，本书的重点在于介绍该领域的基本概念、基本过程和方法，各种算法以介绍其适用条件和原理为主，尽量少涉及具体算法的数学公式。

本书的特点之二是以应用为主。介绍数据挖掘方法在医学研究和服务中的应用实例，为学生今后进一步从事这方面的深入研究提供基础。

本书的特点之三是在内容组织上力求全面系统，突出重点。本书由浅入深、突出交叉学科的特色的同时，注重介绍知识的层次，适合不同水平读者的学习需要。

面对着这样一门还处于蓬勃发展中的学科，对于一群对数据挖掘的理论和技术还没能全面掌握的编者而言，这不啻一场十分艰苦的战斗。好在我们有着对专业的热爱和对事业的追求，所有的编写人员克服了重重困难，终于不辱使命，圆满完成了任务。至于战果如何，我们充满信心，无论如何，这是我们所知道的国内第一部医学数据挖掘教材，我们的努力付出了，相信大家会在字里行间看到我们的智慧和汗水。

我们等待着读者和教学过程的检验，我们期待着您的批评和建议。

崔　雷

2005 年 10 月 27 日于沈阳

郑重声明

目　录

基　础　篇

核 心 篇

应 用 篇

基　礎　篇

第一章

概　述

摘要：作为本教材的开篇，本章概括介绍了数据挖掘和知识发现的轮廓。这一章包括4个部分：知识发现和数据挖掘的定义，知识发现和数据挖掘的算法、工具和步骤，数据挖掘系统的体系结构以及知识发现和数据挖掘的应用。

关键词：数据挖掘　知识发现

Abstract: As the opening chapter of this textbook, this chapter provides an outline of this field. It includes 4 sections: the definitions of Knowledge Discovery in Database (KDD) and Data Mining (DM), the basic algorithm、tools and procedures of KDD and DM, the architecture of DM systems, and the applications of KDD and Data Mining.

Keywords: Data Mining　Knowledge Discovery in Database

第一节　数据挖掘与知识发现的基本概念

一、数据挖掘的产生

（一）数据爆炸和知识贫乏的现状

南美作家和诗人博尔赫斯(Jorge Luis Borges)曾写过一篇名为《巴比塔图书馆》的小故事。故事中描述了一个无限大的图书馆，它是一个由无数个阅览室组成的无边的网络，每个阅览室里装满书架，书架上排满了图书。尽管大多数书没什么意思，并且其标题也是诸如“啊啊 呀呀呀”之类的毫无意义的东西，人们还是在这个图书馆里转来转去直到死。学者们有一个疯狂的假设：“在某处必定有一个总目录”；“人们能想到的所有的图书肯定在图书馆中的某一地方存放着”。这些假说没有一个能够被证实，也就是说，图书馆里有无限的数据但是没有知识。

巴比塔图书馆恰好成为当代人类面临的悲惨困境的写照，我们生活在信息“爆炸”的时代，“淹没”在数据的海洋之中。

一方面，信息的传播媒介的种类大大增加了，如网络、广播、电视、报刊、会议等；传播媒介的数量（如电视台、广播电台、报社等）也大量增加。另一方面，信息量也成倍地增长。如报纸杂志无论是品种数量还是版面数量都较过去大大增加；数据库的种类、形式和各自的规模发展迅猛；网络信息更是难以估量。据估计，全世界目前数据量大致每年翻一番，同时有意义的信息数量却急剧下降，我们面临着太多的数据和太少的知识。

信息量的不断增长造成了人们寻找有意义的数据愈加困难，如何使人们能够快速有效地获取自己所需信息，成为广大信息工作者的重要研究课题。正是这种需求催生了一门目前在信息领域里最为活跃、最令人激动的领域——数据挖掘和知识发现。

（二）知识发现和数据挖掘科学领域的诞生

1989年8月，在美国底特律召开的第11届国际人工智能联合会议(International Joint Conference on Artificial Intelligence, IJCAI)上，专门组织了有关知识发现(Knowledge Discovery in Databases, KDD)的专题

讨论会，首次提出了从数据库中发现知识的概念。随后引起国际人工智能和数据库等领域专家的广泛关注，1991 年麻省理工学院（Massachusetts Institute of Technology，MIT）出版社出版了《Knowledge Discovery in Databases》一书。1995 年，在加拿大蒙特利尔召开了首届数据挖掘与知识发现国际学术会议（KDD'95），这次会议是数据挖掘与知识发现领域里的一次具有里程碑意义的会议。此后，知识发现与数据挖掘国际学术会议每年召开一次。1996 年 MIT 出版社又出版了《Advances in Knowledge Discovery and Data Mining》一书。1997 年，Knowledge Discovery and Data Mining 杂志创刊。

经过十多年的努力，数据挖掘技术的研究已经取得了丰硕的成果，发表了大量的研究论文，并且出现了许多成功的应用案例。不少软件公司已研制出数据挖掘软件产品，并在北美、欧洲等国家得到应用。例如，IBM 公司开发的 QUEST 和 Intelligent Miner；Angoss Software 开发的基于规则和决策树的 Knowledge Seeker，Advanced Software Application 开发的基于人工神经网络的 DB Profile，加拿大 Simon Fraser 大学开发的 DB Minner；SGI 公司开发的 MineSet 等。

随着国外知识发现的兴起，我国也很快跟上了国际步伐。1997 年在新加坡召开了第一次亚太知识发现与数据挖掘会议（Pacific-Asia Conference on Knowledge Discovery and Data Mining，PAKDD），此后每年召开一次，1999 年 4 月在北京召开了第三届亚太知识发现与数据挖掘会议（PAKDD'99）。国内的许多科研单位和高等院校竞相开展知识发现的基础理论及其应用研究，包括清华大学、中科院计算技术研究所、空军第三研究所、海军装备论证中心等。我国各大科研资助项目（如“国家自然科学基金”、“973”、“863”及“攻关”）都设立了 KDD 的研究课题。一些企业也开展了此类项目的研制和开发。《计算机世界报》技术专题版于 1995 年 3 月发表了由中国电子设备系统工程公司研究所李德毅教授组织的 KDD 专题；于 1995 年 4 月发表了由中国科学院史忠植研究员组织的“机器学习、神经网络”专题；于 1995 年 12 月发表了由国防科技大学陈文伟教授组织的“机器发现和机器学习”专题，对我国开展知识发现的研究起到了一定的推动作用。从 1989 年至 2005 年 8 月，维普数据库中的《中文科技期刊数据库》（全文版）中共收录有关知识发现的论文 920 篇，数据挖掘相关文献 4 059 篇，涉及数据挖掘和知识发现的中文文献共计 4 498 篇。

二、什么是数据挖掘和知识发现

（一）数据挖掘和知识发现的定义

数据挖掘（Data Mining），也有人称之为“信息抽取”（Information Extraction）、“数据考古学”（Data Archeology）、“信息收割”（Information Harvesting）、“智能数据分析”（Intelligent Data Analysis）等。从英文字面上理解，挖掘（mine）就是抽取（extract），通常是指从地下抽取隐藏的贵重资源的挖掘操作。把 data 和 mining 联系起来，就是对数据进行深入的研究，目的在于从大量的数据中发现事先没有注意到的额外信息。

数据挖掘的定义比较多，人们从多种角度对这一新兴的技术和研究领域加以概括，比较公认的定义是：在数据中正规地发现有效的、新颖的、潜在有用的，并且最终可以被读懂的模式的过程。

从数据挖掘的定义中可以看出，数据挖掘具有如下几个特点。

（1）挖掘对象是超大型的数据库；

（2）发现隐含的知识；

（3）可以用于增进人类认识的知识；

（4）不是手工完成的。

“在数据库中发现知识”的定义是：从数据集中识别出有效的、新颖的、潜在有用的以及最终可理解的模式的非繁琐过程。

上述知识发现的定义涉及如下几个概念。

（1）数据集：是一个有关事实 F 的集合（如学生档案数据库中有关学生基本情况的记录）。它是用来描述事实的有关方面的信息。一般说来，这些数据都是准确无误的。

(2) 模式:对于集合 F 的数据,可以用语言 L 来描述其中数据的特性。只有当表达式 E 比列举的所有 F_E 中的元素的描述方法更为简单时,才可以称之为模式。如,"如果成绩在 81 ~ 90 之间,则成绩为优良"可以称为表达式,而"如果成绩为 81、82、83、84、85、86、87、88、89 或 90,则成绩为优良"就不能称之为一个模式。

(3) 过程:通常是指知识发现的多阶段处理方式,涉及数据的准备、模式搜索、知识评价以及反复的过程优化。

(4) 非繁琐性:包括两个方面,首先,它是对数据的更深层的处理过程,并不是仅仅对数据进行简单的数学运算或者查询,而是找出隐藏在数据背后的信息;其次,还要有一定程度的智能性、自动性。

(5) 有效性:是指发现的模式对于新的数据仍然保持着一定的可信度(正确程度)。

(6) 新颖性:要求发现的模式应该是新的,至少对于系统来说应该如此。模式新颖性的评判方法有两个:其一,对于得到的数据,将其与以前的数据或者期望得到的数据进行比较,来判断该模式的新颖度;其二,对于其内部所包含的知识,通过对比发现的模式与已有的模式的关系来判断。

(7) 潜在有用性:发现的知识将来有实际效用,如用于临床诊断系统里可以降低医疗卫生服务的费用,提高医疗卫生服务的质量。

(8) 最终可理解性:要求发现的模式能被用户理解,目前它主要体现在简洁性上。

简单地讲,知识发现表示了从低层数据抽象高层知识的整个过程。通过数据库中的知识发现,人们可以从数据库中数据的相关集合中抽象有用的知识、数据的规律性或高层次的信息。

从上面对数据挖掘和知识发现的历史介绍中,我们了解到这两个概念是密切相关的,甚至在有些时候被混淆起来使用。在第一届国际数据挖掘与知识发现大会上,有人提出知识发现用于描述从数据中抽取知识的全过程;数据挖掘则仅用于表示知识发现过程中的发现阶段。也有人认为,知识发现用于表示比从数据中寻找知识更为广泛和"高层次"的概念;而数据挖掘一词则表示给决策人员显示和分析数据的高级技术和工具。目前,大多数的统计学者、数据分析人员和信息管理系统的学者偏爱使用数据挖掘一词,而人工智能和机器学习的研究人员多使用知识发现。数据库作为知识发现和数据挖掘的对象,将在第二章中详细介绍,本章主要阐述两者与机器学习和统计学的关系。

(二) 数据挖掘和知识发现与相关学科领域的关系

数据挖掘与知识发现是一门交叉性学科,涉及机器学习、模式识别、统计学、智能数据库、知识获取、数据可视化、共性能计算、专家系统等领域,内涵极为广泛。它从这些学科中汲取相关的技术和理论,但同时又与这些学科存在差异。通过分析数据挖掘和知识发现与这些相关学科之间的异同,可以从不同角度了解数据挖掘和知识发现的基本属性。从根本上讲,数据挖掘和知识发现主要来源于数据库、统计学和机器学习三大主要技术。

1. 数据挖掘与人工智能和机器学习 学习是生物中枢神经系统的高级整合技能之一,是人类获取知识的重要途径和人类智能的重要标志,而机器学习则是计算机获取知识的重要途径和人工智能的重要标志,是一门研究怎样用计算机来模拟或实现人类学习活动的学科。

机器学习(Machine Learning)的定义是:"如果一个系统能够通过执行某种过程而改变它的性能,这就是机器学习"。从定义中可以看出,首先,学习是一个过程;其次,学习是对一个系统而言;最后,学习改变系统的性能。

人工智能(Artificial Intelligence, AI)是有关计算机智能化的理论和技术的研究。该领域的目标是提高计算机的实用性,同时也提高对人类智能机制的理解。人工智能的研究人员根据认知心理学的原理研究各种机器学习的方法。以符号运算为基础的机器学习代替了以统计为基础的机器学习,成为人工智能研究的主流。

数据挖掘利用了人工智能和机器学习领域的诸多最新成果,同时,这些领域的新兴技术如神经网络和决策树,在数据量和计算能力允许的情况下,可以自动地完成很多重要的功能。数据挖掘把这些高深复杂

的技术封装起来，使人们不必掌握这些技术也可以完成任务，从而能更加专注于自己所要解决的问题。

数据挖掘与机器学习都是从数据中提取知识的技术，其区别在于机器学习主要是针对特定模式的数据进行学习，数据挖掘则是从实际的海量数据源中抽取知识，即数据挖掘必须对 10 万条以上记录的数据集有很好的性能。由于数据挖掘处理的数据量非常巨大，数据的完整性、一致性及正确性都难以保证，因此，数据挖掘的算法需要对大量数据具有适应性，算法的效率、有效性和可扩充性亦很重要。

2. 数据挖掘与统计学 统计学和数据挖掘有着共同的目标，即发现数据中的结构。完成数据挖掘的任务要依靠统计分析方法，就数据挖掘算法本身而言，其中很大一部分可以从数理统计中获得理论的解释。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，因此，其预测的准确度还是令人满意的。

但是，统计分析方法应用到数据挖掘领域之后，数据挖掘的专业人员会更关心数理统计算法应用于大批量数据上的有效性、算法性能的优化以及标准的接口等软件实现上的问题。也就是说，数据挖掘作为一个研究领域，往往从计算机的层面进行全局考虑，即从系统的角度进行分析，因为数据挖掘技术从一开始就是面向应用的，它不仅是面向特定数据库的简单检索查询调用，而且要对这些数据进行微观、中观乃至宏观的统计、分析、综合和推理，以指导实际问题的求解，企图发现事件间的相互关联，甚至利用已有的数据对未来的活动进行预测。一个算法如果只能对几百条数据进行分析，那么再完美也是没有用的。

数据挖掘与统计学的不同之处还表现在数据挖掘必须处理混杂的数据字段，包括数字的多样性、数据类型的多样性等。例如，医学数据中就包含了 SPECT 的图像，心电图信号，诸如体温、胆固醇水平、尿检验数据的临床信息以及医生用自然语言书写的解释。

三、数据挖掘的知识表示

数据挖掘获得知识的表现形式主要有 6 种：规则、决策树、知识基、网络权值、公式和案例。

（一）规则

规则由前提条件和结论两个部分组成。前提条件由字段项（属性）取值的合取（与 $\wedge$）和析取（或 $\vee$）组合而成，结论为决策字段项（属性）的取值或者类别组成。

例如，两类人群的 9 个元组（记录）如表 1 - 1 所示。

表 1 - 1 两类人群数据例

	身高	头发颜色	眼睛颜色
第一类人	矮	金色	蓝色
	高	红色	蓝色
	高	金色	蓝色
	矮	金色	灰色
第二类人	高	金色	黑色
	矮	黑色	蓝色
	高	黑色	蓝色
	高	黑色	灰色
	矮	金色	黑色

如果利用数据挖掘方法，会很快得到如下规则知识：

IF（头发颜色 = 金色 $\vee$ 红色）$\wedge$（眼睛颜色 = 蓝色 $\vee$ 灰色）THEN 第一类人

IF（头发颜色 = 黑色）$\vee$（眼睛 = 黑色）THEN 第二类人

即，凡是具有金色或者红色的头发，并且同时具有蓝色或灰色眼睛的人属于第一类人；凡是具有黑色头发或黑色眼睛的人属于第二类人。

（二）决策树

数据挖掘的信息论方法所获得的知识一般表示为决策树。如 ID3 方法的决策树是由信息量最大的字段(属性)作为根节点,它的各个取值为分枝,对各个分枝所划分的数据元组(记录)子集,重复建树过程,扩展决策树,最后得到相同类别的子集,以该类作为叶节点。

例如,上例的人群数据库,按 ID3 方法得到的决策树如图 1 - 1 所示。

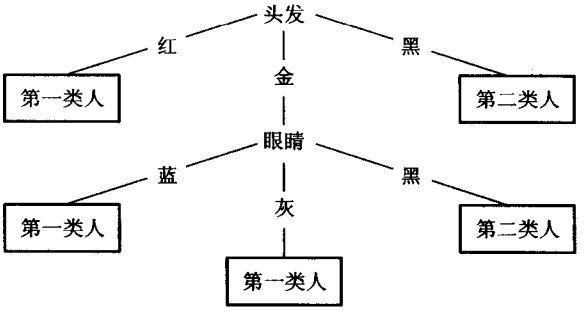


图 1 - 1　决策树

(三) 知识基(浓缩数据)

数据挖掘方法能计算出数据库中字段项(属性)的重要程度,对于不重要的字段可以删除,对数据库中的元组(记录)能按一定的原则合并,这样,通过数据挖掘的方法能大大压缩数据库的元组和字段项,最后得到浓缩数据,称为知识基。它是原数据库的精华,很容易转换为规则知识。

例如,上例的人群数据库,通过计算可以得出身高是不重要的字段,删除该项后,再合并相同的数据元组,得到的浓缩数据如表 1 - 2 所示。

表 1 - 2　知识基(浓缩数据)

	头发	眼睛
第一类人	金色	蓝色
第一类人	红色	蓝色
第一类人	金色	灰色
第二类人	金色	黑色
第二类人	黑色	蓝色
第二类人	黑色	灰色

(四) 网络权值

神经网络方法经过对训练样本的学习之后,所得到的知识是网络连接权值和节点的阈值,一般表示为矩阵和向量。例如,异或问题(如判断某物是苹果还是橘子,但不是两者皆是)的网络权值和阈值分别如图 1 - 2 (a)、(b)所示。

(五) 公式

对于科学和工程数据库,一般存放的是大量的实验数据(数值)。它们中蕴涵着一定的规律性,通过公式发现算法,可以找出各种变量间的相互关系,并用公式表示。

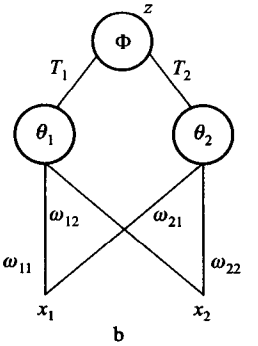
例如,太阳系行星运动数据中包含行星运动周期(旋转一周所需要时间,d)以及它们与太阳的距离(围绕太阳旋转的椭圆轨道的长半轴,百万千米),具体数据如表 1 - 3 所示。

$$\begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{pmatrix} = \begin{pmatrix} 11 \\ 11 \end{pmatrix}$$

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} 0.5 \\ 1.5 \end{pmatrix}$$

$$(T_1, T_2) = (-1, 1)$$

a



b

图 1 - 2　异或问题的网络权值和阈值

表 1-3　太阳系行星数据

	水星	金星	地球	火星	木星	土星
周期 p	88	225	365	687	4 343	10 767
距离 d	58	108	149	228	778	1 430

通过物理定律发现系统 BACON 等工具可以得到开普勒第三定律：$d^3/p^2=25$。

（六）案例

案例是指人们经历过的一次完整事件。当人们要解决一个新问题时，总是先回顾自己以前处理过的类似事件（案例），利用以前案例中解决问题的方法或者处理结果作为参考并进行适当的修改，以解决当前的新问题。利用这种思想建立基于案例推理（Case-Based Reasoning，CBR）。CBR 的基础是案例库，在案例库中存放着大量成功或者失败的案例。CBR 利用相似性检索技术，到案例库中搜索与新问题相似的案例，再经过对旧案例的修改来解决新问题。

可见，案例是解决新问题的一种知识。案例知识一般表示为三元组。

（1）问题描述：对求解的问题及周围世界或环境的所有特征的描述。

（2）解描述：对问题求解方案的描述。

（3）效果描述：描述解决方案后的结果情况，是失败还是成功。

第二节　知识发现和数据挖掘的步骤、算法与工具

一、知识发现和数据挖掘的基本步骤

如前所述，一般认为数据挖掘是知识发现的关键步骤。从技术的层面上讲，知识发现的基本步骤包括数据选择、处理、转换、数据挖掘和解释与评价几个阶段（图 1-3）。

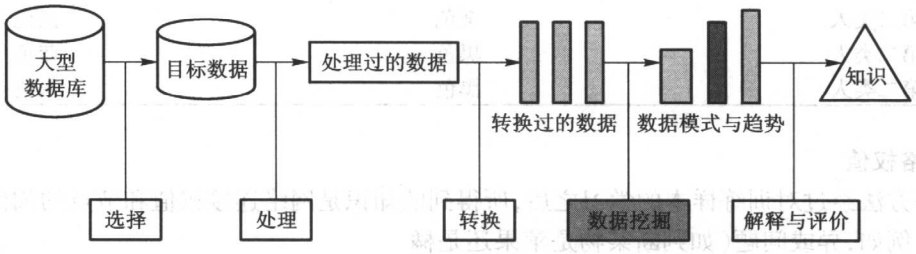


图 1-3　知识发现与数据挖掘的基本步骤

（1）选择：根据某种标准选择或者切分数据。例如，将所有患有肺结核的患者的记录收集起来，形成该疾病患者的数据子集。

（2）处理：包括清除和充实两个方面。由于数据是来自于日常工作中的记录，有许多冗余和重复的内容，如患者的姓名可能在药局和实验室的数据库中都出现；有时还要从其他数据库中补充新数据等。

（3）转换：删除那些丢失重要内容的记录，将数据分类（如按患者的年龄将其分成不同的年龄组），改变记录的格式（如将生日转换为实际年龄）等。

（4）数据挖掘：运用工具和算法，在数据中发现模式和规律，这些具体工具和算法是本书的主要内容。

（5）解释与评价：将发现的模式解释成为可以用于决策的知识，如预测、分类任务、总结数据库的内容