DATA MINING ALGORITHMS AND APPLICATIONS

数据挖掘算法与应用

梁 循

北京大学出版社
PEKING UNIVERSITY PRESS

数据挖掘算法与应用

梁　循

北京大学出版社
PEKING UNIVERSITY PRESS

内 容 提 要

数据挖掘是一个涉及数据库技术、计算智能、统计学、模式识别等多个学科的领域。目前，数据挖掘已经在各行各业有了非常广泛的应用。

本书综合了大量国内外的最新资料和作者的研究成果，系统地介绍了数据挖掘算法、相关技术及其金融数据上的应用。在绪论之后，全书从结构上分为3篇。第1篇具体介绍了数据挖掘的主要算法，包括决策树算法、神经网络算法、基因算法、基本统计分析方法、贝叶斯网络算法、支持向量机方法等。第2篇主要讨论数据挖掘的相关技术，包括数据仓库技术、模糊处理技术、粗糙集技术以及目标优化技术。第3篇探讨了一些数据挖掘的应用专题，包括互联网金融信息搜索引擎、互联网信息流时间序列挖掘等问题。

本书的读者可以是对金融应用感兴趣的计算机专业人士，也可以是对计算机和互联网感兴趣的金融专业人士。它可供数据挖掘、机器智能、金融数据分析等领域的科技人员和高校师生参考。

图书在版编目(CIP)数据

数据挖掘算法与应用/梁循 编著. —北京：北京大学出版社，2006.4
ISBN 7-301-08737-3

Ⅰ.数… Ⅱ.梁… Ⅲ.数据采集 Ⅳ.TP274

中国版本图书馆CIP数据核字(2006)第024257号

书　　名：数据挖掘算法与应用
著作责任者：梁循
责任编辑：沈承凤
标准书号：ISBN 7-301-08737-3/TP·0780
出版发行：北京大学出版社
地　　址：北京市海淀区成府路205号　100871
网　　址：http://cbs.pku.edu.cn
电子信箱：zpup@pup.pku.edu.cn
电　　话：邮购部62752015　市场营销中心62750672　编辑部62752038
排　　版者：兴盛达打字服务社　58745033
印　　刷者：北京宏伟双华印刷有限公司
经　　销者：新华书店
787毫米×1092毫米　16开本　20.5印张　511千字
2006年4月第1版　2006年4月第1次印刷
定　　价：29.00元

前　言

20世纪末，数据挖掘技术在发达国家工业界如雨后春笋般地发展起来，这一方面是由于数据挖掘算法经历了多年的发展，已经到了瓜熟蒂落的阶段；另一方面，也是由于计算机，特别是互联网的普遍使用，使得数字型的数据产生更为方便、传播也更为迅速，造成各专业公司积累了极其大量的数据，远远超过了人类的手工处理能力，于是计算机数据挖掘技术顺理成章地成了各专业公司首要选择。作者当时身处美国硅谷工业界，曾亲手开发过数据挖掘的产品，亲身感受了这种来自应用的需求和推动的巨大力量。

本书的应用部分也涉及一些金融方面的内容，介绍了最近几年一些金融数据挖掘方面的问题。本书是作者的另一本书籍《网络金融》的姊妹书籍，是该书最后一章的扩展。

从结构上，第1章绪论主要回顾和讨论了数据挖掘的概念、算法、应用领域、软件产品及其开发商，其余的部分可以粗略地分为三篇。

第1篇为数据挖掘算法篇，介绍了目前流行的主要数据挖掘算法。第2章决策树算法讨论了构建和修剪决策树的方法、ID3算法、C4.5算法、CART算法、SLIQ算法以及SPRINT算法。第3章神经网络算法主要介绍了单层神经网络和多层神经网络的原理、反向传播算法，分析了多层神经网络算法的机理和算法的改进技术，也简单介绍了径向基函数网络、竞争学习和侧抑制问题、自组织特征图、反馈网络、随机算法和Boltzmann网络，并讨论了神经网络在金融市场中的应用问题。第4章基因算法主要讨论了基因算法的基本步骤、原理和应用。第5章介绍了基本统计分析方法，包括假设检验和区间估计、成组数据的比较、回归分析、方差分析，并列举了基本统计分析方法在互联网股市数据挖掘中的应用例子。第6章专题讨论统计学中的贝叶斯网络方法，主要包括基本概念、beta和Dirichlet分布、贝叶斯网络及其学习问题、不完全数据情形下的学习。第7章主要介绍线性可分问题的支持向量机方法、线性不可分问题的支持向量机方法、核函数、libSVM仿真平台以及支持向量机方法在股市预测上的应用。第8章简要介绍了其他数据挖掘方法，包括主成分分析、近邻法、期望值最大化方法、粗糙集技术、*K*-均值聚类、*K*-中心点算法和关联分析。

第2篇为数据挖掘相关技术篇，主要包括和数据挖掘相关的一些技术。第9章讨论了数据仓库技术，包括数据仓库的结构、设计和关联以及OLAP等。第10章模糊处理技术讨论了特征函数、隶属度函数、截集概念、模型识别、模糊关系以及模糊聚类。第11章粗糙集技术讨论了不可分辨关系、下近似和上近似、近似精度、粗糙集隶属函数以及模糊集与粗糙集的关系。第12章目标优化技术简要讨论了数据挖掘中常用的优化模型以及典型的搜索算法。

第3篇为数据挖掘算法的应用篇。第13章研究的是网络金融中的数据挖掘问题，概述性地探讨了互联网数据挖掘以及网络金融数据挖掘的概念、意义、特点和内容。从整体上看，互联网金融数据挖掘是一个刚刚开始的课题，很多问题还需要研究。比如，体系尚不明朗，包含范畴也不很明确等。第14章讨论的是互联网数据挖掘问题，即互联网金融信息搜索引擎。主要研究了金融定点收割引擎和金融爬虫搜索引擎以及搜索引擎定价问题。第15章将互联网

金融信息流量化，形成时间序列，并从时间序列的角度分析和挖掘了互联网金融信息流的内涵，讨论了互联网金融信息与金融市场从量上的关联，列举、研究和展望了的一系列新的挖掘课题。

目前，各高校普遍更新了或正在更新自己的传统课程，并开设了不少新的课程。这是积极地顺应当前学科中高速的知识更新以及因学科交叉而不断产生出新领域的潮流的结果。同时，也造成了各高校的教学亟需大量新教材和新参考书籍的局面。由于这些领域的成果常常发表在最新的国内外学术期刊上，如果将它们整理，并汇集在一起，就会大大方便教师和学生的使用，提高我国科技工作者的效率。本书所介绍和探讨的数据挖掘处于上述这样一个年轻的交叉领域。

回忆2000年作者在斯坦福大学校园的书店里看到第一本数据挖掘的书的喜悦情景，已经过去了6个年头，因为数据挖掘这个词很准确地概括了作者过去和当时进行的很多看起来散乱的工作，所以，本书也包含了不少作者在科研和工业界多年积累的结果。

作者在本书的编写过程中，得到所在单位北京大学计算机科学技术研究所领导的大力支持以及一些同事和同学的帮助。在此一并致谢。

本书和《网络金融》一起组成北京大学计算机科学技术系研究生选修课教学和辅导材料。课程网页上提供另外一些材料，网页的地址是 http://www.icst.pku.edu.cn/course/efinance2005/index.html。

由于作者水平和时间的限制，书中一定存在不少缺点和错误，恳请读者批评指正。作者也准备将更正及时发表在网上，作为本书的一个补充。

梁 循

2006年2月于北京大学承泽园

目　录

第 1 篇　数据挖掘算法

第2篇 数据挖掘相关技术

第 3 篇　数据挖掘应用

第1章 概 论

1.1 数据挖掘的定义和范畴

数据库量的迅猛增长向科学家、工程师和销售人员提出了一个挑战：在大量的数据中，是不是会隐藏着有价值的东西呢？如何充分有效地使用这些数据进行科学研究、系统优化以及发现商业数据内在关系及其可说明的问题，便成为一个很重要的课题。

为实现这个目的，人们发展了各种各样的算法：统计、机器学习、神经网络、推理网络、决策树以及针对各种特定实际问题的解决方法。数据挖掘涉及的学科领域和方法很多。这门新兴的边缘科学结合了统计学、机器学习、模式识别、智能数据库、知识获取、人工智能、专家系统、数据可视化及高性能计算等领域。它已吸引了计算机科学家、工程技术人员、认知科学家和统计学家的极大兴趣。

随着信息技术的发展，特别是互联网的发展和信息量的爆炸性增长，信息的重要性与日俱增。如何有效地获取有用的互联网信息与知识，是数据挖掘的目标所在；另一方面，互联网为数据挖掘提供了良好的挖掘环境与挖掘对象，且其挖掘结果易于应用，获得直接可见的回报。在这种应用环境与应用需求的刺激下，数据挖掘越来越受到重视。

目前尚无关于数据挖掘的精确学科划定，从广义上来讲，数据挖掘(data mining，DM)先从巨大的数据体系或数据库里提炼出我们感兴趣的东西(可能在我们意料之中，也可能在我们意料之外)，或者说，从庞大的观察数据集中提炼并分析出不能轻易察觉或断言的关系，最后给出一个有用的并可以理解的结论。简单地说，数据挖掘就是在数据中发现模式、知识，或数据间的关系。

数据挖掘也常被称为知识发现(knowledge discovery，KD)，这就无怪乎许多知识发现中的算法，比如人工智能，常常被使用于数据挖掘的过程中。1989 年，在第 11 届国际人工智能的专题研讨会上，学者们首次提出了基于数据挖掘的知识发现(knowledge discovery in database，KDD)概念。1995 年在美国计算机年会上，一些学者开始把数据挖掘视为数据库知识发现的一个基本步骤或把两者视为近义词讨论。

数据挖掘也已经成为数据库理论和应用的一个很重要的方向。事实上，数据挖掘和这些新的数据库研究方向紧密相关，比如数据仓库可看作数据挖掘的一个预处理过程，移动数据库、互联网数据库、海量数据库、并行数据库、空间数据库等新型数据库的挖掘方法是数据挖掘的前沿课题之一。

数据挖掘有如下三个特点：

第一，数据挖掘的数据量常常是巨大的。因此，如何高效率地存取数据，如何根据一定应用领域找出数据关系即高效率算法以及是使用全部数据还是使用一部分随机或有目的地选择出的数据子集，都成为数据挖掘工作者要考虑的问题。

第二，数据挖掘面临的数据常常是为其他目的而收集好的数据(比如说，银行已存有巨大

的每日出入账的数据，这些数据原本是为其他目的而存储的)(梁循，2006)。这就为数据挖掘提出了一个问题，即收集数据时，可能有一个或几个重要的变量未被收集，而这些变量在后来做数据挖掘时被证明是有用的，甚至是至关重要的。也就是说，未知性和不完全性将始终伴随数据挖掘的过程。

第三，数据挖掘的另一个特点是数据挖掘工作者常常不愿把先验知识预先嵌入算法内，因为这样就等于做“假设检验”(但这不排除把统计中的假设检验作为其中间一步来做)。数据挖掘常常要求算法主动性地提示一些数据内在的关系。新颖性是衡量一个数据挖掘算法好坏的一个很重要的标准。当然，这些新颖性的结论必须是可以被人理解的，绝对不应该是漫无边际的奇怪结论。

很显然，数据挖掘有别于传统的数据查询、报表及全文检索等数据分析的方法，它常常是在没有前提假设的情况下，从事信息的挖掘与知识的提取。数据挖掘所得到的信息结果，当然不一定全都是先前未知的。

1.2 数据及其度量

数据集是指从特定场合或过程中测量和记录下来的一组数据。简单地说，就是对目标物体的采集，而对每一个物体，我们采用同一种度量标准。这种度量标准可能是一维的，也可能是多维的，比如对一组病人，度量标准可以是病历号、年龄、性别、症状等。如果我们有 n 个物体，m 个度量标准，那么我们就有了一个 $n\times m$ 的数据集。

数据可以有各种各样的形式，比如有的数据是数字形式的，有的数据是音像形式的。正如我们前边提到的，数据挖掘是一门由许多门学科交叉产生的边缘学科，而这些学科内部本身已发展出了处理这类特定问题的各种有效算法。当然，这些数据最终都可能转换成数字形式的数据，然后加以处理。

数据可以有各种各样的分类，比如连续型数据和离散型数据。连续型数据就是指可以取实数值的那些度量，比如速度、时间、温度这些物理量。离散型数据是取整数值的那些度量，比如每月的天数，某一顾客存取款项的次数，两极的极性。离散型数据的一个特例是二值数据，比如两极的极性。在实际应用中，连续型数据和离散型数据有时没有严格的划分。在某一特定领域(比如银行)内，数据挖掘工作者也把可精确到最细单位的度量(比如币值)看做是连续量处理。

数据常常带有一定的不确定性，或每一个数据都伴随有一个概率，整个度量形成一个概率分布。这些不确定性大都是在数据采集、解译和数据入库等过程中由于仪器设备精度限制和人为因素造成的，很多是不可避免的，即使通过后期数据处理也只能减少空间数据的部分不确定性。不确定性在现实世界是客观存在的，确定性是相对的，不确定性是绝对的。这些不确定性在空间数据挖掘过程中的传播，造成误差的积累。所以，概率方法、模糊方法在很多场合下给出的结论更具有说服力。当然，引入概率在多数情况下导致计算时间增长、算法复杂性增大。应当说明的是，有时这种增加计算时间和算法复杂性的做法在考虑其收益和付出比时是不值得的。这些都是数据设计者应当考虑的问题。

来自现实世界的数据体常常带有污染或误差，有时也叫作噪声，这实际为数据挖掘提出了更多的挑战。在一些情况下，这些污染也可以被认为是白噪声(而事实上，大部分数据挖掘的

算法是基于此假设的，因为这样处理在数学上最方便，结果也最明了，又不失一般性）。

数据挖掘过程中，数据的质量决定了能否获得有意义的分析结果。数据准备阶段大约占了数据挖掘整个工作量的80%。数据准备包括：剔除冗余数据、保证数据的逻辑一致性等。

数据集根据用途可以划分为建模用数据集和检验用数据集。根据不同领域的叫法，建模用数据集也常常被称作样本集、训练数据集，或干脆就叫数据集。检验用数据集常常也称为检验数据集。如何随机地，或按比例地，或按一定算法从庞大的数据集中选出以上这两个数据集合，这也是一个研究的课题。

数据挖掘过程本质上是通过度量数据之间的距离和相似程度来实现的。为了度量数据之间的距离和相似程度，我们需要首先定义一些变量。在此基础上可以定量地进行分析。

设有 n 个样本，组成样本集

$$X = \{X_1, X_2, \cdots, X_n\}$$

其中，样本 i 为一个 m 维向量 $X_i = (x_{i1}, x_{i2}, \cdots, x_{im})^T$，$i=1,\cdots,n$，$(\cdot)^T$ 为$(\cdot)$的转置。每一维分量表示一个属性 a_j，$j=1,\cdots,m$。所有属性组成属性集

$$A = \{a_1, a_2, \cdots, a_m\}$$

也就是说，n 个样本可以视为 m 维属性空间中的 n 个点。

有时，为了方便，我们也使用如下变量属性矩阵

$$B = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nm} \end{bmatrix}$$

其中，b_{ij} 可以是二值的，也可以是取任意整数的。

下面，我们首先定义样本之间的距离和角度，然后根据距离和角度，进一步定义样本之间的相似度和相异度。

1.2.1 距离和角度

设 $d(X_i, X_j)$ 为样本 X_i 和 X_j 之间的距离。一般地，距离函数 $d(X_i, X_j)$ 应满足如下条件：

(1) $d(X_i, X_j)=0$，当且仅当 $X_i=X_j$；

(2) 非负性：$d(X_i, X_j)\geqslant 0$；

(3) 对称性：$d(X_i, X_j)=d(X_j, X_i)$；

(4) 三角不等式：$d(X_i, X_j)\leqslant d(X_i, X_k)+d(X_k, X_j)$。

应用中，常见的有闵可夫斯基距离和马氏距离。

1. 闵可夫斯基距离

闵可夫斯基(Minkowski)距离为

$$d(X_i, X_j) = \| X_i - X_j \|_q = \left(\sum_{k=1}^{m} | x_{ik} - x_{jk} |^q \right)^{\frac{1}{q}}$$

其中，$q\in[1, +\infty]$。

由于从1到$+\infty$，q 都可以任意取值，闵可夫斯基距离是无限个距离度量的概化。特别地，当 $q=1$ 时为曼哈坦距离，当 $q=2$ 时为欧几里得距离，当 $q\to+\infty$ 时为切比雪夫距离。

(1) 曼哈坦距离

曼哈坦(Manhattan)距离为

$$d(X_i,X_j)=\|X_i-X_j\|_1=\sum_{k=1}^{m}|x_{ik}-x_{jk}|$$

(2) 欧几里得距离

欧几里得(Euclid)距离为

$$d(X_i,X_j)=\|X_i-X_j\|_2=\left(\sum_{k=1}^{m}|x_{ik}-x_{jk}|^2\right)^{\frac{1}{2}}$$

(3) 切比雪夫距离

切比雪夫(Chebyshev)距离为

$$d(X_i,X_j)=\|X_i-X_j\|_\infty=\max_{k\in\{1,2,\cdots,m\}}|x_{ik}-x_{jk}|$$

2. 马氏距离

注意到样本各分量的观测值往往为随机变量,所以,第 i 个样本观测值 $X_i=(x_{i1},x_{i2},\cdots,x_{im})^T$ 应为随机向量。随机向量有一定的分布规律,各个分量之间还可能相关,因此,两个样本作为两个随机向量的个体,将第 i 个与第 j 个样本间马氏(Mahalanobis)距离的平方记为

$$d^2(X_i,X_j)=(X_i-X_j)^TA^{-1}(X_i-X_j)$$

其中,A 是随机变量的协方差矩阵。在 A 未知的情况下,也常使用其估计值。

有时,还可以根据每个变量的重要性为其赋一个权重,例如,加权的欧几里得距离形式为

$$d(X_i,X_j)=\|X_i-X_j\|_2=\left(\sum_{k=1}^{m}w_k|x_{ik}-x_{jk}|^2\right)^{\frac{1}{2}}$$

在分析中需要根据数据类型、应用目标等因素选择距离函数。

3. 夹角余弦

设样本 X_i 和 X_j 之间的夹角为 $\theta(X_i,X_j)$。显见,$\theta(X_i,X_j)\in[0,+180]$。夹角余弦为

$$\cos[\theta(X_i,X_j)]=\frac{\sum_{k=1}^{m}x_{ik}x_{jk}}{\left(\sum_{k=1}^{m}x_{ik}^2\sum_{k=1}^{m}x_{jk}^2\right)^{\frac{1}{2}}}$$

1.2.2 相似度和相异度

相似度是度量样本之间相似程度的度量。下面给出相似度和相异度的一般性定义。

如果映射 $s:X\times X\to R$ 满足,$\forall i,j,h$:

(1) $s(X_i,X_j)\geqslant 0$;

(2) $s(X_i,X_j)=s(X_j,X_i)$;

(3) $s(X_i,X_j)\leqslant s(X_h,X_h)$;

(4) $s(X_i,X_i)=s(X_j,X_j)$。

(即自己和自己最相似,并在此时达到最大值)

则称 s 为相似(测)度,或近似系数。

如果映射 $d:X\times X\to R$ 满足,$\forall i,j,h$:

(1) $d(X_i,X_j)\geqslant 0$;

(2) $d(X_i,X_j)=d(X_j,X_i)$;

(3) $d(X_i,X_i)=0$。

(即自己和自己最不相异,或最相似,并在此时达到最小值)

则称 d 为相异(测)度,或相异系数。

显见,夹角余弦是相似测度,向量自己和自己的角度为0时,夹角余弦为1(最大),相似度为1(最相似);夹角越大,夹角余弦越小,相似度越小。距离是相异测度,向量自己和自己的距离为0时,相异度为0(最不相异);距离越大,相异度越大。

1.2.3 类之间的距离

1. 类的定义

设 $X^{(c)}$ 为元素的集合,它共有 $n^{(c)}$ 个元素,记为 $X_i,i=1,2,\cdots,n^{(c)}$。另外,再给定一个阈值 $\theta^{(c)}>0$。

我们知道,类是相似事物的集合。从数学的角度,则难以给出一种通用的严格的定义。常用的有以下几种定义,可以适用于不同的场合。

(1) $\forall X_i, X_j\in X^{(c)}, d(X_i,X_j)\leqslant\theta^{(c)}$,则称 $X^{(c)}$ 为一类;

(2) $\forall X_i\in X^{(c)}, \dfrac{\sum\limits_{\substack{j=1\\ j\neq i}}^{n^{(c)}} d(X_i,X_j)}{n^{(c)}-1}\leqslant\theta^{(c)}$,则称 $X^{(c)}$ 为一类;

(3) $\forall X_i\in X^{(c)}, \exists X_j\in X^{(c)}, j\neq i, d(X_i,X_j)\leqslant\theta^{(c)}$,则称 $X^{(c)}$ 为一类。

显见它们均通过限制元素间的距离来定义类,只是限制的程度不同。第一个定义的要求最高,凡满足它的条件,一定也满足其他定义的条件。

2. 类的特征

我们由以下几种方法来描述类的特征。

(1) 类的重心

类的重心为各元素的均向量

$$\overline{X}^{(c)}=\frac{\sum\limits_{i=1}^{n^{(c)}}X_i}{n^{(c)}}$$

(2) 类的直径

类的直径定义为

$$d(X^{(c)})=\sum_{i=1}^{n^{(c)}}(X_i-\overline{X}^{(c)})^T(X_i-\overline{X}^{(c)})$$

(3) 类的样本离差矩阵与样本协方差矩阵

类的样本离差矩阵为

$$A^{(c)}=\sum_{i=1}^{n^{(c)}}(X_i-\overline{X}^{(c)})(X_i-\overline{X}^{(c)})^T$$

类的样本协方差矩阵为

$$S^{(c)}=\frac{A^{(c)}}{n^{(c)}-1}$$

3. 类间的距离

两个聚类 $X^{(c_1)}$ 和 $X^{(c_2)}$ 之间的距离 $d(X^{(c_2)},X^{(c_2)})$ 的计算方法是分层聚类算法的基础。最常用的距离度量有以下几种。

(1) 最近距离

两个类之间的最近距离由分别来自于它们而且距离最近的两个点决定(见图 1-1):

$$d(X^{(c_1)},X^{(c_2)}) = \min_{X_i \in X^{(c_1)}, X_j \in X^{(c_2)}} d(X_i,X_j)$$

(2) 最远距离

两个类之间的最远距离由分别来自于它们而且距离最远的两个点决定(见图 1-1):

$$d(X^{(c_1)},X^{(c_2)}) = \max_{X_i \in X^{(c_1)}, X_j \in X^{(c_2)}} d(X_i,X_j)$$

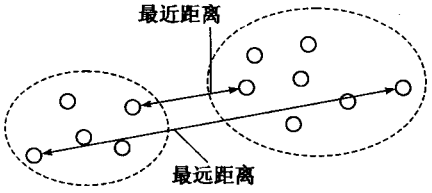

图 1-1 最近距离和最远距离

(3) 均值距离

最近距离度量和最远距离度量代表了类与类之间距离的两个极端,就像所有利用最大值或最小值的算法一样,它们对某些噪声和孤立点都非常敏感,用平均值代替它们显然可以改善这些问题。

两个类之间的均值距离由分别来自于它们的所有点决定:

$$d(X^{(c_1)},X^{(c_2)}) = \frac{\sum_{X_i \in X^{(c_1)}} \sum_{X_i \in X^{(c_2)}} d(X_i,X_j)}{|X^{(c_1)}||X^{(c_2)}|}$$

(4) 基于类直径的距离

基于类直径的距离为

$$d(X^{(c_2)},X^{(c_2)}) = d(X^{(c_2)} \cup X^{(c_2)}) - d(X^{(c_2)}) - d(X^{(c_2)})$$

1.3 数据挖掘的过程

首先,让我们来看看数据挖掘的全过程。一般认为,数据挖掘(真正的生成模式和预测模型的过程)只是将数据转化为知识的过程中的一步。

图 1-2 显示的是从技术的角度看到的数据挖掘过程,它的重点是对数据预处理的优化,略去了真正成功地用于一个商业应用时需要的很多步骤。这种数据挖掘视图表示了如何从原始数据得到有用的模式,再进一步得到知识。数据挖掘工具越好,从一个步骤到另一个步骤的转化就越自动化、越简单。

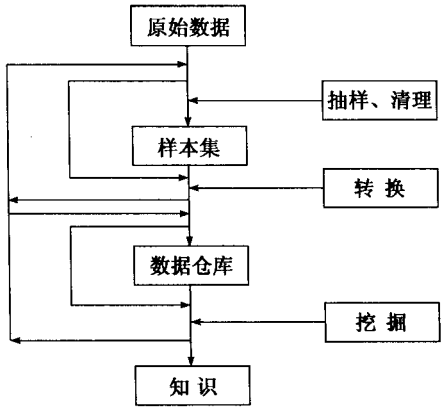

图 1-2 数据挖掘的基本过程

在数据挖掘的过程中，有时需要重复以上的某些步骤。此外，以上四步的分界线常常不好截然划分，比如说，数据预处理及转换本身也可以看作是一种线索关系的提炼。因而这种划分只是旨在说明数据挖掘常常要做的工作。

在开发一个系统的总费用中，原始数据的采集部分占到相当大的比重。当然，采用较小规模的数据对问题的可行性进行初步研究也是可行的，但为了确保将来工作时良好的性能，必须要采集和利用足够多的原始数据。

在数据采集之后，需要进行抽样和清理工作。抽样就是在原始数据中，将有代表性的数据提取出来，这些数据称作样本。清理工作就是将一些不适合用来训练和学习的数据排除在系统之外，这些数据包括不完全数据、噪声数据以及矛盾数据等。

抽样和清理工作的结果就是得到了数据样本集。样本集可以用作训练和学习。当然，如果不想做抽样和清理工作，也可以跳过这一步，直接用原始数据去训练和形成我们的模型。为了保证数据挖掘结果的价值，必须了解数据。输入数据库中的异常数据、不相关的字段或互相冲突的字段、数据的编码方式等都会对数据挖掘输出结果的质量产生影响。虽然一些数据挖掘算法自身会对上面提到的这些问题做一些考虑，但让算法自己做所有这些决定是不明智的。在进行数据挖掘前要对数据进行必要的“整理”与“筛选”，能够提高数据挖掘的效率与正确性。

数据有可能不是我们需要的形式。在这种情形下，我们需要进行数据转换。

转换完成后，如果不满意，就应该返回到上一阶段，从原始数据重新进行抽样工作，即从头开始。如果满意，则可以继续进行下一步。

数据仓库是一种数据存储的有效形式，非常利于数据挖掘（事实上，数据仓库本身也提供了不少数据挖掘的功能）。

数据仓库之后，我们就可以使用各种数据挖掘的算法了。数据挖掘的过程包括特征选择、模型选择、训练和评价等。特征选择就是，根据特定的问题领域的性质，选择有明显区分意义的特征。这常常是设计过程中非常关键的一步。实实在在拿到的样本数据，有利于选择特征。当然，先验知识在其中也具有非常重要的作用。有时，这种知识的嵌入过程可以更微妙或更复杂。在其他一些应用中，知识或许来源于被考查的模式的形态和它的特定属性。

以分类问题为例，在选择或设计特征的过程中，很显然，我们希望发现那些容易提取、对不

相关变形保持不变、对噪声不敏感以及对区分不同类别的模式很有效的特征集。但是，怎么做才能把先验知识和实验数据有机结合起来，以发现有用的和有效的特征，是一个技术加艺术的问题。在模型选择中，我们需要首先知道设定的类别模型，并努力保证它与真实世界的模型一致。大体地说，利用样本数据来确定分类器的过程称为训练分类器。在设计模式识别系统时，没有一个通用方法可以解决所有的问题。一般认为，反复试验和“基于样本的学习”的方法是设计分类器最有效的方法。

使用数据挖掘算法的结果，应该产生我们所需要的知识。但是，如果我们没有得到所需要的知识，则需要返回到上一阶段，甚至最开始，重新执行上述过程。

有时，从理论上说，数据挖掘算法确实可以得到准确的结果，例如分类器可以达到无错误的分类结果，但是这时对处理时间和存储容量的要求都惊人的大，甚至根本无法实现。因此，考虑不同算法的计算资源消耗和计算复杂度有着重要的实践意义。比如说有些问题中，我们知道在不考虑工程上的约束的前提下，确实能够设计一个性能非常优秀的分类器。但是如果存在工程上的约束，情况就不完全一样甚至完全不一样了。因为前者是在实验室里完成的，而后者要在现场环境下工作。

1.4 数据挖掘的任务和建模

目前，随着信息化技术的快速发展，数据信息的收集和整理都极为便捷。人们面对的已不是局限于本部门、本单位或本行业的数据库，而是浩瀚无垠的信息海洋。数据挖掘常常应用于含有大量计算机或仪表、能自动直接产生巨大的数字型数据资源的场合，比如医药、金融、电子工程等领域。人们积累的数据越来越多，信息过量几乎成为人人需要面对的问题。数据的丰富带来了对高效的数据分析方法的需求。现代计算机技术与数据库技术可以支持具有大容量存储和快速查询功能的数据库，但对于这种“整齐有序”却“堆积如山”的数据集合，传统的分析方法已经很难适应。

数据挖掘能够发现的知识可以分为：广义型知识、特征型知识、差异型知识、关联型知识、偏离型知识和预测型知识等。广义型知识是反映同类事物共同性质的知识；特征型知识是反映事物各方面的特征知识；差异型知识是反映不同事物之间属性差别的知识；关联型知识是反映事物之间依赖或关联的知识；偏离型知识揭示事物偏离常规的异常现象；预测型知识根据历史的和当前的数据推测未来数据。

分类和聚类是数据挖掘中两项最常见的任务。分类是数据挖掘中一项非常重要的任务，目前在商业上应用最多。分类的目的是提出一个分类函数或分类模型，该模型能把数据库中的数据项映射到给定类别中的某一个。分类和回归都可用于预测。预测的目的是从历史数据记录中自动推导出对给定数据的推广描述，从而能对未来数据进行预测。和回归方法不同的是，分类的输出是离散的类别值，而回归的输出则是连续数值。而聚类是根据数据的不同特征，将其划分为不同的数据类。它的目的是使得属于同一类别的个体之间的距离尽可能的小，而不同类别上的个体间的距离尽可能的大。聚类方法包括统计方法、机器学习方法、神经网络方法和面向数据库的方法。

数据挖掘中还常见到相关性分析的内容。相关性分析的目的是发现特征之间或数据之间的相互依赖关系。数据相关性关系代表一类重要的可发现的知识，这类知识可被其他模式抽

取算法使用。一个依赖关系存在于两个元素之间。如果从一个元素 A 的值可以推出另一个元素 B 的值，则称 B 依赖于 A。这里所谓元素可以是字段，也可以是字段间的关系。数据依赖关系有广泛的应用。依赖关系分析的结果有时可以直接提供给终端用户。然而，通常强的依赖关系反映的是固有的领域结构而不是什么新的或有兴趣的事物。自动地查找依赖关系可能是一种有用的方法，常用技术有回归分析、关联规则、贝叶斯网络等。

偏差分析或称孤立点分析也是数据挖掘的主要任务之一。偏差分析包括分类中的反常实例、例外模式、观测结果对期望值的偏离以及量值随时间的变化等，基本思想是寻找观察结果与参照量之间的有意义的差别。异常包括如下几种可能引起人们兴趣的模式：不满足常规类的异常例子、出现在其他模式边缘的奇异点、在不同时刻发生了显著变化的某个元素或集合、观察值与模型推测出的期望值之间有显著差异的事例等。

数据挖掘比较常见的成功应用包括：解释性数据的分析、描述性建模、预测性建模、知识发现、模式和规则的识别、文字图像信息的搜寻等。

顾名思义，解释性数据分析是指，在进行数据挖掘之前我们没有明确的想法或思路，常常是交互式的方式，通过观察来分析数据。对较小规模的度量个数少的数据集，很多软件公司开发出了可以提供视觉分析的图形表达工具。这些工具对有效地进行交互式数据挖掘很有帮助。随着度量个数增多，视觉表达方法变得越来越难于应用。

描述性建模就是对数据及产生数据的过程进行描述。用户常常需要抽象的有意义的描述。经过归纳的抽象描述能概括大量的关于类的信息。有两种典型的描述：特征描述和判别描述。特征描述是从与学习任务相关的一组数据中提取出关于这些数据的特征式，这些特征式表达了该数据集的总体特征；而判别描述则描述了两个或更多个类之间有何差异。

在描述性建模的典型方法中有的对数据的整体概率分布进行描述，用聚类的方法将数据体分成若干组，即进行分组和组分析。描述式数据挖掘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。有的方法主要分析数据内部各度量变量之间的关系。概念描述就是对某类对象的内涵进行描述，并概括这类对象的有关特征。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，后者描述不同类对象之间的区别。例如，可以研究去年销售额增加 50%的产品的特征。举例子来说，在做市场分析时，隔离和分组时常常把相似的数据分成一组。这样就把购买特性相同的顾客群挑了出来，从而有利于对他们的性别、年龄、购买个性进行分析。一个众所周知的例子是在超市中凡是购买婴儿尿布的人常常也买啤酒。商家进而分类对他们做广告或提出折扣内容。当然，具体分多少组没有规定或算法，这要根据实际情况而定。

预测性建模的目的是根据现有数据集及已知度量去推测求知度量，比如股市在未来某一时刻的高低，或者一个产品不易测度到的变量(除非弄坏它)。预测式数据挖掘是通过建立一个或一组模型，对数据进行总结、聚类、关联规则发现、序列模式发现、依赖关系或依赖模型发现、异常和趋势发现工作等，试图预测新数据集的行为。预测性建模在很多领域里已得到长足发展，其中大部分是建立在统计学和人工智能方法上的成功预测应用。数据挖掘代替传统的手工分析问题方法，自动在数据库中寻找出预测性信息，具有高效、准确和时效性特征，能自动预测趋势和行为。

知识发现的任务之一是对数据进行总结。总结的目的是对数据进行浓缩，给出它的紧凑描述。传统的也是最简单的数据总结方法是计算出数据库的各个字段上的求和值、平均值、方