数据挖掘技术及其应用
Data Mining Techniques and Its Applications

刘同明　等编著

国防工业出版社

·北京·

图书在版编目(CIP)数据

数据挖掘技术及其应用/刘同明等编著 .—北京:国防工业出版社,2001.9

ISBN 7-118-02504-6

Ⅰ.数... Ⅱ.刘... Ⅲ.数据处理 Ⅳ.TP274

中国版本图书馆 CIP 数据核字(2001)第 12836 号

国防工业出版社出版发行

(北京市海淀区紫竹院南路 23 号)

(邮政编码 100044)

北京奥隆印刷厂印刷

新华书店经售

*

开本 850×1168 1/32 印张 8¼ 211 千字

2001 年 9 月第 1 版 2001 年 9 月北京第 1 次印刷

印数:1—2000 册 定价:19.00 元

前　言

为解决知识获取这一困扰人工智能的瓶颈问题，20 世纪 80 年代末提出了知识发现或在数据库中发现知识(KDD)。1995 年提出了数据挖掘(data mining)概念，作为知识发现过程的关键步骤。数据挖掘的目的是，把人工智能、机器学习与数据库等技术结合起来，由计算机自动从已有数据(数据库或数据仓库)中发现以前未知的、具有潜在应用价值的信息或模式，解决数据量很大、而知识贫乏的矛盾。这一概念一经提出，立即引起学者、软件开发商和用户的极大兴趣，国外纷纷建立了许多专门研究知识发现或数据挖掘的公司或部门，从数据挖掘的基本概念和原理开始，直至挖掘方法、算法以及软件工具，进行了广泛深入的研究。到目前为止，已经形成了比较完整的数据挖掘理论和方法体系，并且出现了许多实用的数据挖掘工具，广泛用于商业、金融、保险、医疗、化工、制造业、工程和科学等领域，产生了巨大的效益。

数据挖掘是数据库技术、人工智能、机器学习等多种学科相结合的产物，本书第一章概要介绍了数据挖掘的一般概念和原理。数据挖掘的主要技术方法基础有：归纳学习、神经网络、遗传算法以及最近邻方法等，本书第二、三、四、五章，分别详细介绍最为热门的面向属性归纳学习、基于粗糙集理论以及神经网络技术和遗传算法的数据挖掘方法。分布式数据挖掘和空间数据挖掘等是今后非常重要的研究方向，本书第六章对这些方面都有较好的论述。第七章结合应用和开发，介绍数据挖掘方法和工具的选择，以及作者开发的一个辐射源特征识别数据挖掘专用系统。

数据挖掘是一种理论性和应用性都很强的技术。作者通过大量例子介绍数据挖掘有关的概念、方法和算法。数据挖掘又是一

个年轻而充满活力的研究领域，技术发展很快，因此，作者在各章中尽可能面向未来，提出一些研究方向和热点，供读者参考。

全书共分七章，第一、二、六、七章由刘同明撰写，第三、四、五章分别由房靖、吴小俊、曹奇英执笔，最后由刘同明统稿。解洪成研究员和夏祖勋教授通读了全书，提出了许多宝贵的建议和意见。刘伟为本书提供了大量参考文献，曾富贵、谈丰编写了部分章节，并实现了辐射源特征挖掘系统，周天才等做了许多文字编辑和绘图工作。作者所在单位华东船舶工业学院领导也非常关心本书的写作，提供了各种条件。因此，本书是集体智慧和努力的结晶。

东南大学校长顾冠群院士和南京大学博士生导师蔡士杰教授在百忙之中为本书撰写了推荐意见；中国舰船研究院侯正明教授和总装备部系统工程所何新贵教授，以及国防图书基金委员会的专家教授，也给了我们很多关心和指导。在此，我们致以衷心的感谢！同时，也感谢国防工业出版社的大力支持！

作者的本意是抛砖引玉，但限于才疏学浅，加之数据挖掘是一项新技术，书中难免有不当之处，敬请专家、学者和读者批评指正！

作　者

第一章 绪 论

数据挖掘是20世纪90年代中期兴起的一项新技术，它是知识发现过程中的关键步骤。国内外学术界和企业界，都非常重视对数据挖掘技术和软件工具的研究和开发。

数据挖掘是多门学科和多种技术相结合的产物，也是一个非常年轻而又活跃的研究领域。本章概要介绍数据挖掘的概念、任务、方法、原理、应用领域，以及今后的研究方向。

第一节 知识发现和数据挖掘

面对信息社会中数据和数据库的爆炸式增长，人类分析数据和从中提取有用信息的能力，远远不能满足实际需要。虽然数据库管理系统(DBMS)可以高效实现数据录入、检索和维护等管理功能，但不能发现数据中的关联和规则，也不能根据现有的数据预测未来的发展趋势。所以，迫切需要一种能够智能地自动地把数据转换成有用信息和知识的技术和工具。需求是发展之母，数据库管理系统和人工智能中机器学习两种技术的发展和结合，促成了在数据库中发现知识(KDD)这一新技术的诞生。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的专题讨论会上，首次提出KDD。它是一门交叉性学科，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、高性能计算、专家系统等领域，内涵极为广泛，理论和技术难度很大，从而使针对大型数据库的KDD技术一时还难以满足应用需要。于是，1995年的(美)计算机学会(ACM)会议提出了数据挖掘(data mining)

概念，它形象地把大型数据库看成是存放有价值信息的矿藏，通过有效的知识发现技术，从中挖掘或开采出有用的信息。

所谓数据挖掘，就是从数据库中抽取隐含的、以前未知的、具有潜在应用价值的信息的过程[1]。也有一些文献把数据挖掘称为知识抽取（knowledge extraction）、数据考古学（data archaeology）、数据捕捞（data dredging），等等[2]。多数人认为数据挖掘是KDD过程中的关键步骤（见图 1－1），从而不加区分地使用知识发现和数据挖掘这两个术语。

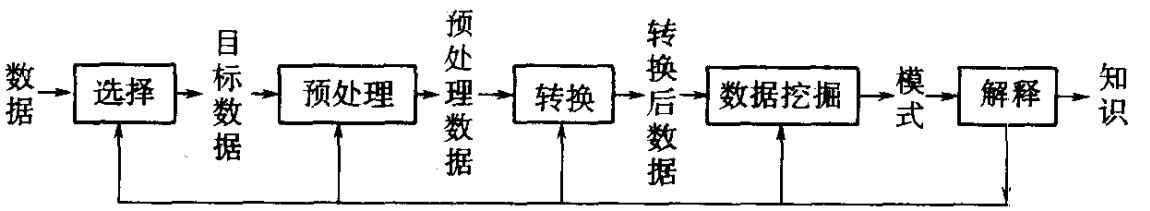

图 1－1　KDD过程

数据挖掘与传统数据分析工具的主要区别在于它们探索数据关系时所使用的方法。传统数分析工具使用基于验证的方法，即用户首先对特定的数据关系作出假设，然后使用分析工具去确认或否定这些假设。这种方法的有效性受到许多因素的限制，如提出的问题和预先假设是否合适等。与分析工具相反，数据挖掘使用基于发现的方法，运用模式匹配和其他算法决定数据之间的重要联系。

第二节　数据挖掘的任务

一般把知识表示成规则形式。按照数据挖掘技术所能够发现的规则，将常见的数据挖掘任务分为七种类型：

① 特征规则挖掘。特征规则是一个断言，它把由所有数据满足的概念特征化。特征规则挖掘能够总结并发现由用户指定的数据集的一般特征，如特定疾病的症状等。

② 辨识规则挖掘。发现把一个数据集(目标类)与另一个数据集(对比类)区分开来的特性或性质。例如,为了把一种疾病与另一种疾病区分开,辨识规则总结区分这些疾病的症状。

③ 互联规则挖掘。互联规则描述对象集之间的关联关系,例如,对$\{A_i\}_{i=1}^{m}$和$\{B_j\}_{j=1}^{n}$,可能有形式 $A_1 \wedge \ldots \wedge A_m \rightarrow B_1 \wedge \ldots \wedge B_n$ 的规则。

④ 分类规则挖掘。把被分类数据映射到一组已知的类。例如,根据汽车的汽油 - 里程把汽车加以分类。

⑤ 数据聚类。根据对象属性标识对象集的聚类(类或组)。对象按某种聚类准则聚类后,对象组内的相异性最小,组间的相异性最大。例如,根据疾病症状,把一组疾病聚类成几个类。

⑥ 预测。预测某些被丢失数据的可能值或数据集中某些属性值的分布。例如,根据公司员工的工资分布预测某个员工的工资。

⑦ 趋势性规则挖掘。发现反映数据集中普遍演变行为的规则集。例如,发现影响库存商品价格的因素。

当然,上述七种规则仅仅是目前已知的规则知识中的一部分,尚有其他一些规则类这里未列出,例如量化规则等。

第三节　数据挖掘的分类

对数据挖掘进行分类可以依据所挖掘的数据源类型、所发现的知识类型以及所使用的技术等。

被挖掘的数据源通常是数据库或数据仓库。数据库的数据模型可以是关系的、网状的、层次的或面向对象的,但由于关系数据库的一系列独特优点,而成为当前数据挖掘的主要实施对象,称之为关系数据挖掘,这是本书讨论的主要内容。从应用角度看,被挖掘的数据库有事务数据库、多媒体数据库、分布式数据库、空间数据库,等等,本书第六章将简要介绍其中的分布式数据挖掘和空间

数据挖掘。

也可以根据所发现的知识的抽象层次来分类数据挖掘。知识层次分为通用知识、初级知识和多层知识。一个灵活的数据挖掘系统可以发现多抽象层上的知识。

还可以根据数据挖掘的基础技术来分类。例如，按照驱动方法把数据挖掘分为自主数据挖掘、数据驱动挖掘、查询驱动挖掘以及交互式数据挖掘等[3]。

此外，在统计学、集合论、逻辑学、信息论、认知论、人工智能等学科理论的基础上，提出了各种各样的数据挖掘方法和技术，并且形成了各自的特点和应用领域。但从构成大多数数据挖掘系统的基础来看，实际上只有几种基本技术。

用户的挖掘活动主要有三类：模式发现、预测建模和分析评价[4]。“模式发现”是寻找隐藏在数据集中的模式的过程；“预测建模”则是利用所发现的模式预测未来；“分析评价”就是应用所得到的模式发现异常现象的过程。

预测建模和分析评价都需要把当前输入的新数据与以往的数据集进行匹配，因此需要保存过去的数据。而模式发现则不需要，一旦提取模式后，就可以把过去的数据移去。所以，数据挖掘方法总体上分为两大类：基于数据保持类方法和基于模式提取类方法，如图 1－2 所示。

基于模式提取的数据挖掘方法又分为三种：基于逻辑、基于十字表方法以及基于方程方法，它们各自建立在相应的理论基础上，如逻辑学、集合论或神经网络理论等。一般来说，基于逻辑的数据挖掘方法既能处理数字型数据，也能处理非数字型数据；基于方程方法的挖掘方法则要求所有待挖掘的数据都是数字型的，神经网络是这类方法的代表；十字表方法则相反，它只能处理非数值型数据，典型方法是遗传算法。

现代数据挖掘方法主要依靠模式提取技术，同时，为了改善和提高数据挖掘的功能、性能和效率，发展趋势是综合采用多种方法和技术。

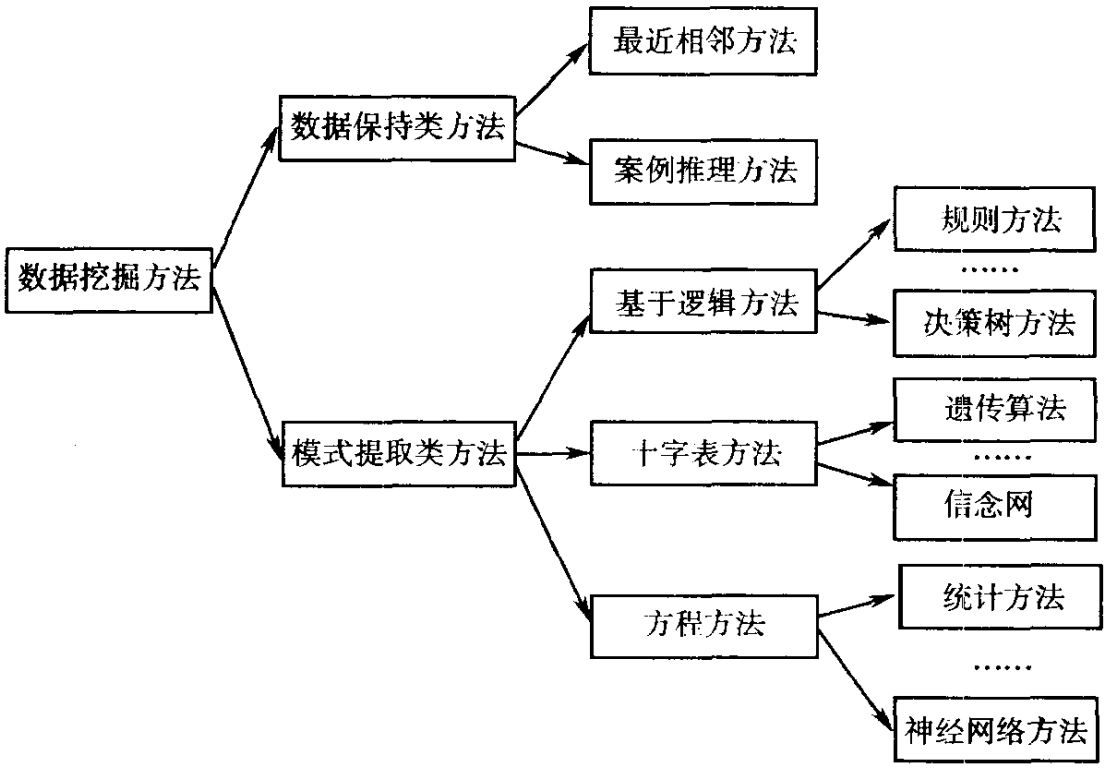

图 1-2　数据挖掘方法和技术分类

第四节　关系数据挖掘的基本原理

关系数据库因为具有坚实的数学基础、统一的组织结构、完整的规范化理论、一体化的查询语言等优点，成为当前数据挖掘的主要对象。

从数据库学习是一种特殊环境(数据库)下的机器学习，所以许多概念和原理与机器学习相同或相似，但又有自己的特点。有关机器学习的概念、原理、方法和技术，请读者参阅人工智能和知识工程方面的书籍。

一、有监督学习和无监督学习

数据挖掘的目的是发现数据集中的模式或规律，这里有两种方式：有监督(教师)学习和无监督(教师)学习。数据挖掘中具有

有监督学习特点的是分类器(classifier),分类器能够根据某种分类准则对输入数据进行分类。教师的作用体现在,系统建立模型时由教师提供一定的帮助,如预先定义类别,提供属于这些类别的正例和反例等。而无监督学习方式下教师不预先定义类别,系统必须自己寻找把对象聚类到类别中去的某种方法,以及对这些类别的描述。数据挖掘中广泛使用有监督学习方式。但是,当对模式缺少先验知识时,还是要使用无监督式数据挖掘方法。

二、训练集和测试集

把数据挖掘系统试图从中提取知识的数据集称为训练集。为了测试所发现的知识的正确性和有效性,用另一个称为测试集的数据来测试。显然,如果在训练集中所发现的模式是正确的,那它对测试数据也应该是正确的;如果从训练集所得到的知识是通用的,那它对绝大部分测试集也是有效的。

三、演绎和归纳

演绎和归纳是两种广泛使用的逻辑推理过程,从认知学和知识建立角度看,大多数数据挖掘软件用归纳法发现知识,而在评价所发现的知识时要用演绎法。由图 1 - 1 可见,从数据库抽取模式的算法是归纳与演绎的结合。

四、泛化和特化

归纳学习的一般操作是泛化(generalization)和特化(specialization)。将训练数据集分为不相交的正例集合和反例集合,正例集合用于泛化,反例集合用于特化。泛化操作用于扩展一假设的语义信息,使其能够包含更多的正例,应用于更多的情况。而特化是泛化的相反操作,用于限制概念描述的应用范围。然而,关系数据库并不明显地存放正例数据和反例数据,也就没有明显的反例数据可用于特化,所以,关系数据库中的归纳学习过程主要依赖于泛化。

五、示例(样本)学习

示例学习(learning from examples)是用于数据库知识发现的重要策略。它是一个从训练样本集表示的多个特定实例归纳出一般概念或规则的过程,常用一个四元组 $<P,N,C,\Lambda>$ 表示,其中,P 是正例集合,N 是反例集合,C 是定义学习任务的概念集合,Λ 是获得特定逻辑结构的运算符集合。

从训练样本集可能归纳出多个结论,为解决这种多样性,使用与数据挖掘任务有关的领域知识(称为背景知识)约束可能的结论空间。概念和逻辑运算符是作为背景知识出现的,以便把所需要的结论限制为带有特定词汇和逻辑公式的概念,学习过程只考虑这些概念。通过归纳产生覆盖所有正例排除所有反例的概念。所以,有些文献把这种学习方法称为覆盖正例排斥反例方法。

控制示例学习过程的策略有数据驱动、模型驱动以及它们的混合,常采用的学习算法有候选消除算法、AQ 算法以及 ID3 算法等[5]。

采用示例学习的关系数据挖掘广泛使用面向属性的归纳学习方法(见第二章)。

六、知识发现任务的描述和说明

描述和说明一个知识发现任务需要:与任务相关的数据;背景知识;所期望的知识表示方法;语言工具。

1. 与发现过程相关的数据

数据库的大量数据中,一般只有一部分与特定的学习任务有关,并且可能涉及到几个关系。为此,用查询方法从数据库中汇集与任务相关的数据,并把这些数据视为学习的例子或样本,因而可以运用示例学习策略。

2. 背景知识

数据质量和广泛性是数据挖掘的关键。常使用背景知识对数据进行清理、精练和补充,以便提高数据质量。为了保证数据的广泛性,使用各种技术对数据进行聚焦,即只把与任务相关的数据汇

集到数据集中。如果要智能地完成数据聚焦，也需要背景知识。

背景知识常用概念层次表示，并用以控制泛化过程。知识发现把概念视为思维的单位，并将不同层次的概念组织成分类目录，按照“从一般到特殊”的次序排列。最一般的概念用保留字“ANY”描述，最特殊的概念对应于数据库的属性值。通过概念层次，把学习到的规则用普遍化的概念表示出来，按大多数用户所希望的简洁形式加以描述。

概念层次可以由知识工程师或领域专家提供，也可以根据数据库中的数据自动或半自动地构造它们(见第二章)。

3. 学习结果的表示

根据人工智能理论，知识可以用命题逻辑、一阶谓词逻辑、产生式、语义网络等方法表示。而神经网络既是推导信息的机制，又是表示知识的方法。但从逻辑角度讲，关系中的每个元组就是一个合取范式的逻辑公式，因此，一个关系可以用这些合取范式的析取来表征。这样一来，已发现的知识也可以表示成关系形式。

4. 语言工具

常见的用于描述和说明数据挖掘任务的语言工具是类 SQL(结构化查询语言)，使用这种语言既可以描述挖掘任务，又可以利用语言的查询功能从数据集中汇集与挖掘任务相关的数据。

第五节　数据挖掘的典型应用领域

数据挖掘能够自动发现以前未知的模式，自动预测未来趋势和行为。因此，数据挖掘技术广泛用于以下一些领域：

① 零售/市场营销。这是数据挖掘技术应用最早也是最重要的领域，主要功能是：市场定位，消费者分析，预测销售趋势，优化营销策略，分析库存需求，选择零售点，价格分析等。在民航系统中，还可以帮助优化组合航线航班，发现提高航线效益的机票预订方式。

② 金融。预测存/贷款趋势，优化存/贷款策略；抽取预测模式；监督交易活动，发现交易规则。

③ 信用保险。分析保险客户的要求和信誉，保险风险行为模式以及欺诈行为。

④ 过程控制/质量监督。鉴别产品制造过程中的缺陷；管理由异常行为引起的通信网络数据。

⑤ 化工/医药。从各种文献资料中自动提取有关化学反应的信息，发现新的有用的化学成分，分析和解释有利于提高产品质量、改进产品功能和增加公司利润的重要数据。

⑥ 工程与科学数据分析。分析科学数据；数据库模式集成；传感数据分析和处理。

⑦ 司法。帮助调查案件，诈骗监测，洗钱认证，犯罪组织分析等。

⑧ 军事信息系统中的目标特征提取、态势关联规则挖掘等。

第六节　数据挖掘的研究方向

数据挖掘是一个年轻而又非常活跃的研究领域，目前面临的问题，除了基础理论和技术方面的外，更重要的是开发和应用。

数据挖掘中经常遇到的技术难题有：

① 超量数据。数据库中数据量的迅速增长，是促进数据挖掘技术发展的原因之一，也是数据挖掘技术首先要解决的问题。以往的大多数机器学习算法都是针对少量(如几百个)样本数据开发的，因而不能用于比它大几千倍的数据库。现有的许多数据挖掘技术的时间复杂度和模式复杂度对数据量的大小也非常敏感[6]。可能的解决方法是尽可能降低算法复杂度，或者使用数据汇集方法限制搜索空间，以及使用比较快的计算机等。

② 噪声数据。通常把数据输入或汇集过程中发生的非系统误差称为噪声。它有两种形式：

第一种是腐烂的属性值。训练样本的某些值相对于它们应有

的值发生了变化，从而使数据模糊不清，这可能与已经建立的规则相冲突。腐烂值还可能改变已经识别到的正确模式。

第二种是遗漏的属性值。相对于训练样本或待识别的对象，可能遗漏一个或几个属性值，从而使数据不完整。如果属性值是在训练集中遗漏的，则系统可能从总体上把这个对象忽略掉，或者作为“未知”对象处理，从而影响规则的准确性。如果一个属性值仅仅在被分类的某个对象上遗漏掉，则系统可能要检测所有匹配规则，并计算最可能的分类，从而降低了响应速度。

虽然数据库管理系统可以通过函数依赖机制在一定程度上消除或减少数据输入中的误差，但实际上的属性值仍然可能是不精确或错误的，从而影响抽取的模式的准确性，造成最终结果的不确定性。发现和表示带噪声数据的模式要用概率方法。

③ 空值。由于数据不可利用或人类本身的知识不完整，导致数据库中个别或某些记录的属性域出现空值现象，或者对某一发现任务来说，不存在其必需的数据记录。空值不仅表示一个未知的值，还说明该属性值是不可利用的。空值现象尤其在关系数据库中经常发生。这种情况给发现、评估和解释模式带来了困难。这时，知识发模型应当具有近似决策能力，方法主要有模糊集合论和粗糙集理论等。

④ 冗余数据。与不完整数据相反，给定的数据集中可能含有冗余的或者不重要的属性，从而增加时间空间开销和结果规则的复杂度。解决这个问题的方法主要是属性约简(见第三章)。

⑤ 动态数据。数据库的基本特点是，库中的内容是动态改变的，它对于发现方法的影响主要表现在：首先，如果发现是作为数据库应用实现的，那么，要求发现方法高效率，并能充分利用DBMS的检索功能。其次，对于动态变化数据，知识发现方法应当具有增量式(或渐进式)学习的能力。

数据挖掘今后可能的研究方向是：

① 加强应用研究，针对不同数据挖掘任务的专用数据挖掘系统。不同的应用领域可能使用多种类型数据和数据库，知识发现

系统应当能够对不同类型的数据和数据库进行有效的数据挖掘。虽然大多数数据库是关系型的，但许多实际应用的关系数据库还可能含有复杂的数据类型，如结构数据和复杂数据对象、超文本和多媒体数据、空间和时态数据、事务数据以及历史数据等，因此，一个功能很强的数据挖掘系统应能对各种复杂数据类型进行挖掘。鉴于数据类型的差异和不同的数据挖掘目的，要求一个数据挖掘系统能够处理各种类型数据是不现实的，应当根据特定类型数据的挖掘任务构造专用的数据挖掘系统，如关系数据库挖掘，空间数据库挖掘等。

② 高效率挖掘算法。为能从大量数据（特别是大型数据库）中有效地抽取信息，数据挖掘算法必须是高效的，即算法的运行时间必须是可预测的和可接受的，带有指数甚至中阶多项式的算法，没有实际使用价值。

③ 提高数据挖掘结果的有效性、确定性和可表达性。已发现的知识应能准确地描述数据库中的内容，并能用于实际领域。对有缺陷的数据应当根据不确定性度量，以近似规则或定量规则形式表示出来。数据挖掘系统还应能很好地处理和抑制噪声数据和不希望的数据。所以要研究度量知识质量的方法，包括模型和工具的实用性和可靠性。

④ 数据挖掘结果的可视化。数据挖掘的最后阶段是分析已发现的知识，有多种不同的分析观点和表示形式，要求用高级语言（最好是自然语言）或图形界面表示数据挖掘需求和已发现的知识，并采用合适的知识表示技术。这样，数据挖掘任务就可由非领域专家指定，也便于用户理解和直接应用已发现的知识。

⑤ 多抽象层上的交互式数据挖掘。要预测从数据库中实际上能够发现什么是很困难的，所以要采用高级数据查询去调查和揭示某些可进一步探索的踪迹。交互式数据挖掘允许用户交互地精练数据挖掘需求，动态改变数据焦点，逐步深化数据挖掘过程，从不同角度不同抽象层次上灵活地观察数据和挖掘结果。

⑥ 多源数据挖掘。计算机网络（包括局域网和广域网，特别

是 Internet)把许多数据源联接在一起,形成巨大的分布式异构数据库。不同来源数据的格式和语义可能不统一,数据挖掘系统应当能够帮助用户揭示异构数据库中的高级数据规律。数据库规模的扩大,数据分布的广泛性,以及分布式数据挖掘的复杂性,促进了并行和分布式数据挖掘算法的研究。在分布式数据挖掘研究方面,今后应特别重视把数据挖掘技术与 Internet 技术及 Web 技术紧密结合起来,开发出基于 Internet 和 Web 的数据挖掘软件工具。同时,加强对分布式软代理(Agent)技术的研究和应用。

⑦ 数据挖掘的安全性和保密性。当从不同角度和不同抽象层次观察数据的时候,就要考虑数据的安全性和保密性,防止侵犯别人隐私和泄漏敏感信息。

⑧ 实现与现有数据库系统或数据仓库的无缝集成,进一步扩大数据挖掘工具的应用范围和提高现有数据的利用率。

第二章　面向属性的归纳学习技术

归纳学习是一种非常重要的数据挖掘方法，许多关于人工智能和知识工程的书籍中都有介绍[8,9]。但由于数据库中的数据量往往很大，影响了归纳效率，需要采取有效措施进行数据约简。面向属性的数据泛化和归纳学习技术是解决这一问题的有效途径。

本章第一节介绍概念层次的基本概念，第二节介绍概念层次的自动生成，第三节介绍概念层次的实现技术，第四节讨论面向属性(AOI)的归纳学习技术。作为表示背景知识的重要工具——概念层次结构，是所有数据挖掘的基础，而 AOI 技术经改进和扩充后可用于面向对象数据库的数据挖掘。

第一节　概念层次

概念层次是表示数据挖掘中背景知识的重要手段，它与 AOI 方法相结合，使 AOI 成为数据挖掘中最为有用的技术。此外，也广泛用于特征规则挖掘、多层知识挖掘、分类和预测等。

一、概念层次

1. 基本概念

概念层次结构(concept hierarchy)表示把一组较低级概念映射到与它们相对应的较高级概念的次序，这种映射可以按偏序关系($\prec$)来组织概念集。$\prec$反映了概念之间的特殊－一般关系，可以用树、格或有向无循环图等来表示，通称为层次结构。如果用树结构表示概念层次，则树结构的所有术语都可以用于概念层次。