数据挖掘的基本过程及方法

朱 琳

(西昌学院 生化系，四川 西昌 615022)

【摘 要】知识发现与数据挖掘是人工智能、机器学习与数据库技术相结合的产物。随着科学数据大量积累和各种数据库的普遍使用，人们又逐步认识到海量数据的利用十分困难、效率低下，而且从中难以获得有价值的指导性意见。另一方面，更多带规律性的信息和知识又往往被大量原始数据淹没。在这种情况下，数据挖掘技术就应运而生，出现在众多学科领域，成为一种强大的化学信息技术。

【关键词】知识发现/数据挖掘；处理过程；聚类；神经网络方法；遗传算法；粗集法
【中图分类号】S652 【文献标识码】A 【文章编号】1673-1891(2005)03-0073-04

1 概述

知识发现与数据挖掘是人工智能、机器学习与数据库技术相结合的产物。在众多以实践为基础的学科中，其理论的发展往往落后于实践。通常只有一小部分知识可以通过理论推导或计算求得，大部分知识只能以“记忆”方式存储起来。这些数据散布在浩若烟海的各类出版物、数据库和网络中，使得信息搜寻困难，获取有用信息更难。随着数据大量积累和各种数据库的普遍使用，人们又逐步认识到海量数据的利用十分困难、效率低下，而且从中难以获得有价值的指导性意见。另一方面，更多带规律性的信息和知识又往往被大量原始数据淹没。在这种情况下，数据挖掘技术就应运而生，出现在学科领域，成为一种强大的信息技术。

2 KDD/DM 的定义

1996年，Fayyad、Piatetsky-shapiror和Smyth将数据库中的知识发现(KDD)过程定义为：从大量的数据中提取有效模式的非平凡过程，该模式是新颖的、可信的、有效的、可能有用的和最终可以了解的。数据挖掘(DM)是KDD过程中的一个特定步骤，它是采用专门算法从大量数据中抽取模式(patterns)的过程。

3 KDD的处理过程

KDD的处理过程如图1.1所示。从图中可见，KDD过程是由多个步骤相互连接、反复进行人机交互的动态过程。

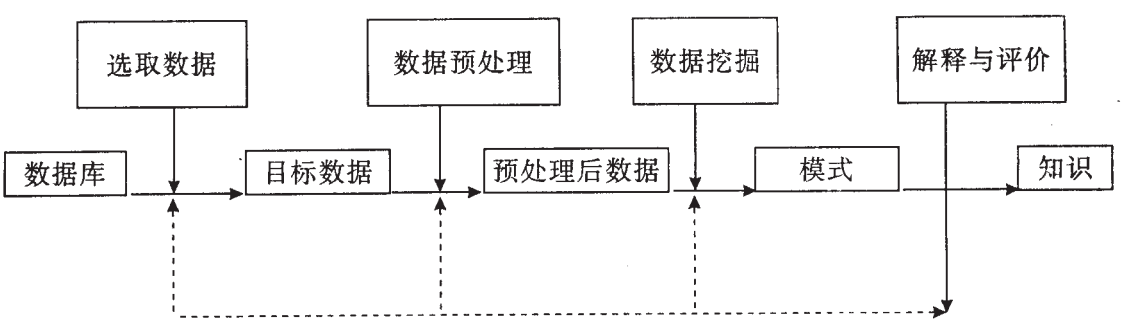

图1.1 KDD/DM的处理过程

实现KDD的具体处理过程包括：

(1)准备：了解KDD相关领域，熟悉掌握有关背景知识，弄清用户要求；

(2)数据选择：根据用户要求从数据库中提取与

收稿日期：2005-06-02
作者简介：朱 琳(1978-)，女，助教。

KDD相关的数据，KDD将主要从这些数据中提取知识，在此过程中会利用一些数据库操作来处理数据，并形成目标数据；

(3)数据清理和预处理：主要对步骤(2)产生的数据进行再加工，检查数据完整性及一致性，处理其中的噪音数据，利用统计方法填补丢失的数据，并考虑时间顺序和数据变化等；

(4)数据缩减：对经过预处理的数据要根据知识发现的任务对数据进行再处理，主要通过换算和投影，找到数据的特征表示，用维变换或转换方法减少有效变量的数目或找到数据的不变式；

(5)确定KDD的目标：根据用户要求确定KDD发现何种类型的知识，对于不同要求的KDD将在具体知识发现过程中采用不同的知识发现算法。

(6)确定用于知识发现的算法：根据步骤(3)所确定的任务选择合适的知识发现算法，这包括选取合适的模型和参数，使知识发现算法与整个KDD的评价标准相一致；

(7)聚焦：即从发掘数据库里选择数据。聚焦方法主要利用聚类分析和判别分析。

(8)数据挖掘(DM)：运用选定的知识发现算法，从数据中搜索或提取用户感兴趣的模式或特定的数据集(即知识)，这些知识可用常规方式表示，如产生式和规则等；

(9)模式解释：对发现的模式进行解释，去掉多余的不切题意的模式，转换某个有用模式，以便于用户理解。

(10)知识评价：将这些知识结合到实际运行系统中，获得这些知识的作用或证明这些知识。用预先和可信的知识检查并解决知识中可能存在的矛盾。

从上述介绍可以看出，数据挖掘是KDD过程中最核心的部分，是采用机器学习和统计等方法学习知识的阶段。与数据挖掘相关的技术之间的关系可用图1.2表示[3]。

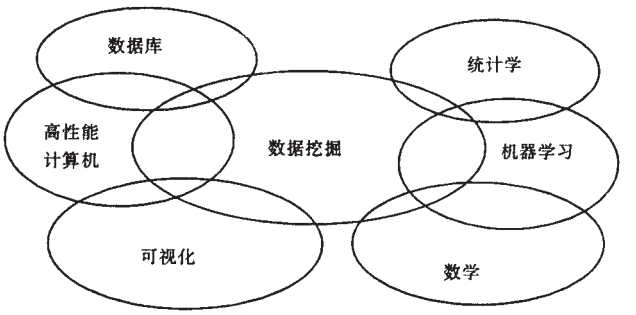

图1.2 数据挖掘的有关技术

4 数据挖掘的分类[2]

数据挖掘涉及的学科领域和方法很多，因此分类的方法也有多种。

按挖掘对象分：有关数据库、面向对象数据库、空间数据库、时态数据库、文本数据源、多媒体数据库、异质数据库、遗产数据库和万维网(WEB)等。

按挖掘方法分：粗略分为机器学习方法、统计学方法、神经网络方法和数据库方法等。机器学习可细分为归纳分析(决策树和规则归纳等)、基于范例学习、遗传算法等。统计方法可细分为回归分析(多元回归、自回归等)、判别分析(贝叶斯判别、费歇尔判别和非参数判别等)、聚类分析(系统聚类和动态聚类等)、探索性分析(主成分分析法和相关分析法)等；神经网络可细分为前馈式神经网络(BP算法)、自组织神经网络(自组织特征映射、竞争学习等)等。

按挖掘任务分：可分为关联规则发现、分类、聚类、时间序列预测模型发现和序贯模式发现等。表1.1列出常用数据挖掘功能、算法及其典型应用领域。

表1.1 常用数据挖掘功能、算法及其典型应用

数据挖掘功能	算法	典型应用领域
关联规则	统计学、集合理论	市场货篮分析、市场分析
分类	决策树、神经网络、粗集	产品营销、定量控制、危险评估
聚类	神经网络、统计学	市场分析
时间序列预测	统计学、ARMA模型	销售预测、利润预测
序贯模式	统计学、结合理论	市场货篮分析

5　数据挖掘常用方法及技术

下面是常用数据挖掘方法与技术的基本要点。

5.1　统计分析方法

主要用于完成总结知识和关联知识挖掘。统计分析方法利用统计学、概率论的原理对关系中各属性进行统计分析,以找出它们之间的关系和规律。统计分析方法是最基本的数据挖掘技术之一。在数据库中,表的属性之间一般存在两种关系:

(1)函数关系 :即能用函数公式表示的、确定性的解析关系;

(2)相关关系 :即不能用函数公式表示,但仍然存在相关的确定关系。常用的统计分析方法有[5]:

常用的统计分析方法有 :判别分析、因子分析、相关分析、回归分析、偏最小二乘回归(PLS)、聚类法(Clustering)等。

聚类分析是数据挖掘中最重要的技术之一。与分类有所不同, 分类的类别是按应用要求事先给定的,根据表示事物特征的数据,可以识别其类别。而聚类的类型不是人为指定的而是分析数据的结果。聚类法大致可分为两种类型:

a. 分层聚类 :分层聚类是基于数学的标准,对数据进行细分或聚合。这种类型适用与数值数据。

b. 概念聚类 :概念聚类基于数据的非数值属性数据进行细分或聚合。这种类型适用与非数值数据类型。

5.2　神经网络方法(Neural Networks)

神经网络方法用于分类、聚类、特征挖掘、预测和模式识别。神经网络方法模仿动物的脑神经元结构,以M-P模型[6]和Hebb学习规则为基础。在本质上是一个分布式矩阵结构,通过对训练数据的挖掘,逐步计算(包括反复迭代或累加计算)神经网络连接的权值。神经网络模型大致可分为以下三种:

(1)前馈式网络 :以感知机、反向传播模型和函数型网络为代表,主要用于预测和模式识别等领域;

(2)反馈式网络 :以Hopfield(人名)离散模型和连续模型为代表,主要用于联想记忆和优化计算;

(3)自组织网络 :以自适应共振理论[7](Adaptive Resonance Theory, ART)模型和Kohonen(人名)模型为代表,主要用于聚类分析。

目前,在数据挖掘中最常用的神经网络是BP网络。当然,人工神经网络还是正在发展的科学,某些理论尚未真正形成,如收敛性、稳定性、局部最小值和参数调整问题等。对于BP网络常遇到的问题是训练速度慢,可能陷入局部最小,以及网络参数和训练参数难以确定等[4]。针对这些问题有人采用人工神经网络与遗传基因算法相结合的办法, 取得了较好的成果[8] 。

人工神经网络具有分布式存储信息、并行处理信息、推理、以及自组织学习等特点,并且具有对非线形数据快速拟合能力, 解决了诸多其它方法难以解决的问题。

5.3　粗集(Rough Set)方法[10]

用于数据简化 (如删除与任务无关的记录或字段)、数据意义评估、对象相似或差异性分析、因果关系及范式挖掘等。Rough集理论是Z.Pawlak在80年代提出来的, 用于研究非精确性和不确定性知识的表达、学习、归纳等方法的。主要思想如下 :在数据库中将行元素看成对象,列元素是属性,把对象的属性分为条件属性和决策属性, 按各属性值是否相同划分等价类。等价关系R定义为不同对象在某个 (或几个)属性上取值相同,这些满足等价关系的对象组成的集合称为该等价关系R的等价类。条件属性上的等价类E与决策属性上的等价类Y之间有三种情况:

(1) 下近似 :Y包含E ;

(2) 上近似 :Y和E的交为空;

(3) 无关 :Y和E 的交为空。对下近似建立确定性规则,对上近似建立不确定性规则(含可信度),对无关情况不存在规则。

5.4　覆盖正例、排斥反例方法

它是利用覆盖所有正例、排斥所有反例的思想来寻找规则。比较典型的有Michalski的AQ11方法、洪家荣改进的AQ15方法和洪家荣的AE5方法。AQ系列的核心算法是,在正例集中任选一个种子,到反例集中逐个比较, 对字段取值构成的选择子相容则舍去,相斥则保留。按这种思想循环所有正例种子将得到正例集的规则(选择子的合取式)。AE系列方法是用扩张矩阵来完成。

5.5　公式发现

公式发现是在工程和科学数据库 (由试验数据组成)中对若干数据项(变量)进行一定的数学运算,以求得相应的数学公式。例如,典型的BACON发现系统就完成了物理学中大量定律的重新发现。它的基本思想是对数据项进行初等数学运算(加、减、乘、除等),形成组合数据项,若它的值为常数项,就得到了组合数据项等于常数的公式。国防科技大学研制

的FDD发现系统，其基本思想是对两个数据项交替取初等函数后，与另一数据项的线形组合若为直线时，就找到了数据项(变量)的初等函数的线性组合公式。该系统所发现的公式比BACON系统发现的公式更为广泛、范围更宽。

5.6 模糊论方法、

利用模糊集合理论对实际问题进行模糊评判、模糊决策、模糊模式识别和模糊聚类分析。模糊性是客观存在的，系统的复杂性越高，精确化的能力就越低，意味着模糊性越强。这是Zadeh总结出的互克性原理。以上提到的模糊方法都已经在化学研究领域取得了较好效果[4]。

5.7 可视化技术[10]

可视化是计算机应用技术的发展趋势，也是数据挖掘的研究方向之一。可视化数据分析技术拓宽了传统的图表功能，用直观图形形式将信息模式、数据关联或趋势呈现给决策者，使之能交互分析数据关系，如把数据库中多维数据变成多种图形对揭示数据总体状况、内在本质及规律至关重要。可视化技术将人的观察力和智能融入挖掘系统，极大提升了系统挖掘的速度、层次和内容。

6 数据挖掘面临的问题

数据挖掘技术虽然在较大范围内得到应用，并取得显著成效，但仍存在着一些尚未解决的问题比如：尚无好方法快速去除或修改噪音数据及处理空缺数据；存在平台支持的局限性；在算法执行过程中，只考虑算法本身的复杂度，缺乏对所利用硬件环境资源的考虑，使得算法实际执行时间长 数据可视化方面还停留在对结果的简单图形描述上；证实技术目前还不太成熟等等。不过相信不久的将来，这些问题必将得到解决。

运用数据挖掘工具，可从大量现有数据库中发掘众多有价值信息，为科学研究提供重要信息，也能大大减少试验的次数，以达到花较少的设计、合成时间，获得预期效果。可以预计，随着计算机网络技术的发展，数据挖掘必然会在将来的研究中发挥更大的作用。

致谢 本文的撰写得到李道华教授的指导，特此致谢！

参考文献：

〔1〕陈文伟，邓苏，张维明.数据开采与知识发现综述.KDD论坛.
〔2〕杨炳儒，江亚东.基于大型数据库的KDD系统及应用研究.科学前沿与学术评论：23(1)49~50.
〔3〕魏长华，王淑礼.知识发现和数据挖掘的研究.高等函授学报(自然科学版)：1999，4(2)44~46l
〔4〕Feng Jiansheng KDD and its applications， Bao Gang techniques，1999(3) 27~31.
〔5〕Guan li，Liang Hongjun Data warehouse and datamining.Microcomputer Applications. 1999，15(9) 17~20.
〔6〕张作生.神经元与神经网络模型.神经网络及其应用.中国科技大学出版社，合肥.1992，58~59.
〔7〕Chen Rong. BP arithmetic and its structure optimization tactics. Journal of Automatization. 1997，23(1)，43~49.
〔8〕郑泽芝 数据库应用新技术—知识发现(KDD) 山西统计，2000(10).
〔9〕唐常杰，杨富华等.数据采掘的基本方法及其专家系统的差异.KDD论坛.
〔10〕吕安民，林宗坚，李成名.数据挖掘和知识发现的技术方法.测绘科学.2000，25(4)36~38.

The Basic Process and Method of Chemical Data Mining

ZHU Lin

(Department of chemistry of Xichang College, Xichang 615022, Sichuan)

Abstract： Along with the large accumulation data and wide using of all kinds of databases， people realize it is very difficult and low efficiency about use of great capacity data and can't get the guidable idea from that. On the other hand， more regular information and knowledge often submerged by large original data. Accordingly， the data mining technology comes into many domains. It is a large information technology and is becoming a new study direction about computer field.

Key words：Knowledge Discovery in Database/Data Mining; The Process; Clustering; Neural Networks; Genetic Algorithms; Rough set