文章编号:1000 - 582X(2004)03 - 0021 - 07

现代数据挖掘技术研究进展*

梁协雄,雷汝焕,曹长修
(重庆大学 自动化学院,重庆 400030)

摘 要:数据挖掘是一个多学科交叉融合而形成的新兴的学科。笔者介绍了数据挖掘的一些基本知识及有关概念,阐述了数据挖掘的一些基本方法(传统的统计学方法、神经网络、决策树、进化式程序设计、基于事例的推理方法、遗产算法、非线性回归方法),然后对当前数据挖掘在各种领域的应用进行了概括,并提出了一些难点(数据质量、信息可视化、极大数据库、信息分析员技能)和今后的研究方向。

关键词:知识发现; 数据挖掘; 数据仓库; 决策支持
中图分类号:TP392 **文献标识码**:A

随着网络技术的发展和计算机的广泛应用,数据化越来越成为一种潮流。但是,人们正面临"数据丰富而知识贫乏"的问题。80 年代末兴起的数据挖掘(data mining)技术或数据库中的知识发现(knowledge discovery in database,KDD)技术为解决此问题开辟了一条道路。数据挖掘是在大量的数据中发现潜在的、有价值的模式和数据间关系(知识)的过程。目前数据挖掘研究和开发表明数据挖掘需要覆盖各种各样不同的应用任务,从数据的预处理到关联规则、聚类分析、数据分类、偏差检查、序列模式等等特定的模式。因此,这一技术的应用是一个极富挑战性的任务。近年来出现的数据挖掘技术之所以被目前认为具有令人兴奋的研究前景,是因为它能够获得广泛的应用[1]。如用于支持企业关键性决策,市场策略的制定、金融欺诈的检测、生物制药等。面对汹涌而来的大量数据,企业对数据挖掘应用形成了极大的需求,将使这一技术迅速得到发展和完善。国外,在大型商业、金融业、保险业、民航等大型企业都开始得到应用。国内目前总体上处于理论及其方法方面的探讨、应用试验阶段[2-3]。

1 数据挖掘的基本内容

1.1 知识发现和数据挖掘

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上首次提出了知识发现 KDD(Knowledge Discovery in Database)这个术语。知识发现(Knowledge Discovery in Database)是指识别出存在于数据库(或数据仓库)中有效的、新颖的、具有潜在效用的、最终可以理解的、模式的、非平凡的过程。也就是说,知识发现是从数据库中发现知识的全部过程。

数据挖掘(Data Mining)是指从大型数据库(或数据仓库)中提取人们感兴趣的知识,这些知识是隐含的、事先未知的、潜在有用的信息,提取的知识一般可以表达为概念(Concepts)[2,4-6]。规则(Rules)、规律(Regularities)、模式(Patterns)等形式。

知识发现的目的是从数据中发现知识,而数据挖掘则是知识发现中的一个特定步骤。二者都是从数据中发现知识,它们的区别可以这样来理解:知识发现比数据挖掘更广泛,而数据挖掘则是更具体、更深入的概念。但在很多地方,就用数据挖掘表示知识发现。

* 收稿日期:2003 - 10 - 15
基金项目:重庆市科委应用基础项目(7369);国家教育部博士点基金项目(98061117)
作者简介:梁协雄(1952 -),男,广东南海人,重庆大学博士研究生。主要从事数据库中的知识发现(KDD)及人工智能系统,数据仓库与数据挖掘的可视化方向研究。

1.2 知识发现的过程

知识发现的过程包括:领域知识的理解、数据的理解、数据的集成与选择、数据预处理、数据挖掘、结果的表达与解释、评价数据挖掘模型、应用所建的模型等步骤,并反复进行人机交互的复杂过程,如图 1 所示。[7-10]。

由此可以看出,知识发现就是应用一系列技术,从大量的、不完全的、有噪声的、模糊的、随机的实际数据中,提取出隐含在数据中的人们实际不知道的但又是潜在有用的信息和知识。数据挖掘算法对数据有一定的要求,如数据冗余性小,数据属性之间相关性小,数据出错率小等,为此数据挖掘必须经过数据准备阶段,以提高数据挖掘的质量;挖掘操作包括选择合适的算法,进行挖掘知识的操作,最后证实发现的知识;表达和解释是对结果进行分析,提出最有价值的信息,如果获得的知识不是决策者满意的,则需要重复以上数据挖掘阶段;最后是把整个知识运用到生产实际中,评价所建立的模型是否能够满足生产实际需要,如果不能,则重新进行整个数据挖掘过程。

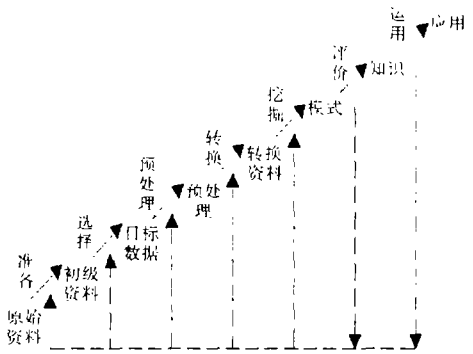

图 1 知识发现的过程

1.3 数据挖掘系统的组成[11-12]

数据挖掘系统由一组构件联合组成,如图 2 所示。

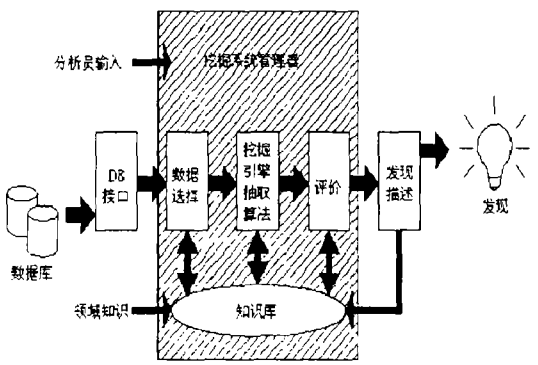

图 2 数据挖掘系统

1.4 数据仓库与数据挖掘[5,13-16]

数据挖掘的对象不仅是数据库,也可以是文件系统或组织在一起的数据集合,然而更主要的是数据仓库。数据仓库是面向主题的、稳定的、不同时间的数据集合,用以支持经营管理中的决策制定过程。通常由一些小型数据库构成,它的主体是由关联数据库构成,但某些层次的数据也可能由其它类型的数据(如多维数据)组成,它兼备数据集成和数据分析的功能。它的作用相当于一个档案室,尽管它的内容允许增加,但一般不允许更新。而数据库中的数据挖掘是一种数据处理过程,它通过机器学习、统计分析或其它方法,从大量数据中提取出有用的信息。数据挖掘和数据仓库协同工作,则可以简化数据挖掘过程中的某些步骤,从而大大提高数据挖掘的工作效率。并且,因为数据仓库中的数据来源很广,因而保证了数据挖掘中的数据来源的广泛性和完整性。

1.5 数据挖掘的特点[11,17-19]

与传统的信息处理方法相比,数据挖掘有其自身的特点:

1) 处理对象为大规模数据库,数据规模十分巨大。

2) 由决策者(用户)提出实时随机查询,靠数据挖掘技术找寻其可能感兴趣的东西。

3) 对于那些因为并没有实际发生或很少发生的行为,它所造成的影响并没有在数据库中体现出来,数据挖掘技术要能够提取出有用的规则,并提出预测。

4) 数据挖掘技术要能快速对数据变化做出反映,并提供决策支持。

5) 数据挖掘既要发现潜在的规则,还要管理和维护规则。

6) 发现的规则不必适用于所有的数据,当达到某一阀值时,便可认为有此规律。

7) 数据挖掘都是根据数据库或数据仓库中的历史资料提取规则,然后用于指导现在以及将来的行动。随着时间的进展和新数据的不断加入,建模所用的数据与当前情况的吻合程度可能降低,这时需要重复建模的过程。

2 数据挖掘的方法

2.1 传统主观导向系统[19-21]

这是针对专业领域应用的系统。如基于技术分析方法对金融市场进行分析。采用的方法从简单的走向分析直到基于高深数学基础的分形理论和谱分析。这种技术需要有经验模型为前提。

传统统计分析:这类技术包括相关分析、回归分析及因子分析等。一般先由用户提供假设,再由系统利用数据进行验证。缺点是需经培训后才能使用,同时在数据探索过程中,用户需要重复进行一系列操作。

以上两种技术主要基于传统的数理统计等数学的基础，适用于数据分析方面。

2.2 神经元网络(NN)[22]

用于分类、聚类、特征采掘、预测和模式识别。神经网络模仿生物神经网络，本质上是一个分布式矩阵结构，它通过对训练数据的采掘逐步计算网络连接的权值。神经网络可分为以下 3 种：

1) 前向多层神经网络[23]

用三层前向多层神经网络可以实现各种非线形影射，其功能加权系数的递推公式如下：

$$\Delta_p W_{ij}^{(3)} = \alpha 2(d_{pi} - O_{pi}^{(3)}) O_{pi}^{(3)}(1 - O_{pi}^{(3)}) O_{pj}^{(2)} \quad (1)$$

其中，$i = 0,1,\ldots,M-1, j = 0,1,\ldots,K$

$$\Delta_p W_{ij}^{(2)} = \alpha\{\sum_{K=0}^{M-1} \delta_{pk}^{(3)} W_{ki}^{(3)}\} O_{pi}^{(2)}(1 - O_{pi}^{(2)}) O_{pj}^{(1)} \quad (2)$$

其中，$i = 0,1,\ldots,K-1, j = 0,1,\ldots,J$

$$\Delta_p W_{ij}^{(1)} = \alpha\{\sum_{K=1}^{k} \delta_{pk}^{(2)} W_{ki}^{(2)}\} O_{pi}^{(1)}(1 - O_{pi}^{(1)}) O_{pj}^{(0)} \quad (3)$$

其中，$i = 0,1,\ldots,J-1, j = 0,1,\ldots,N$

$$\text{注意：} O_{pi}^{(3)} = y_{pi}, i = 0,1,\ldots,M-1 \quad (4)$$

$$O_{pj}^{(0)} = x_{pj}, j = 0,1,\ldots,N \quad (5)$$

本算法的特点是从第三层向前逆推，故称为 BP 算法。由于 S 形变换函数的可微性，给出上列很漂亮的解析结果，考虑到系统的惯性。

$$\Delta_p W_{ij}^{(l)}(t_K) = \alpha\delta_{pi}^{(l)}(t_k) O_{pj}^{(l-1)}(t_k) + \eta\Delta_p W_{ij}^{(l)}(t_{k-1}) \quad (6)$$

通常选择 η 在 0.9 左右。基本 *BP* 算法的缺点是存在局部极值问题，可以用模拟退火、遗传算法及径向基函数等方法加以克服，前向多层神经网络是目前应用最广泛的一种神经网络。

2) 反馈式网络

它以 Hopfield 的离散模型和连续模型为代表，分别用于联想记忆和优化计算；

在图 3 中，每个放大器就是一个神经元，其输出用一个非线性动态方程描述。若第 i 个放大器的输入为 u_i，那么非线性动态方程如下：

$$C_i \frac{du_i}{dt} = \sum_{j=1}^{n} w_{ij} V_j - \frac{u_i}{R_i} + I_i \quad (7)$$

放大器即神经元的输出 V_i 满足：

$$V_i = f_i(u_i) \quad (8)$$

其中，C_i 是放大器的输入电容。w_{ij} 是第 j 个放大器输出到第 i 个放大器输入的联接权，$f_i(\bullet)$ 是第 i 个放大器的输出特性，即神经元特性，这里假设了每个神经元的输出响应时间是可忽略的。n 为神经元的个数，I_i 为放大器的外部输入。

由于 n 个神经元的输出与上式一起可独立地描述 Hopfield 网络的运行状态，因此，把这 n 个输出组成的向量作为系统的状态向量，第 i 输出即为状态向量第 i 个元素。现在状态空间中考虑网络的运行。

为了描述 Hopfield 网络的稳定性，引入如下的 Lyapunov 函数，又称能量函数：

$$E = -\frac{1}{2}\sum_i \sum_j w_{ij} v_i v_j + \sum_i \frac{11}{\tau_i}\int_0^{v_i} f_i^{-1}(v)\,dv - \sum_i I_i v_i \quad (8)$$

在高增益的情况下，上式的第二项可以忽略。

考虑到权重的对称性，可求得 E 的时间导数值为：

$$\frac{dE}{dt} = -\sum_i \frac{dv_i}{dt}\left(\sum_j w_{ij} v_j - \frac{u_i}{R_i} + I_i\right) \quad (9)$$

再根据上式，即有

$$\frac{dE}{dt} = -\sum_i C_i \frac{df_i^{-1}(v_i)}{dv_i}\left(\frac{dv_i}{dt}\right) \quad (10)$$

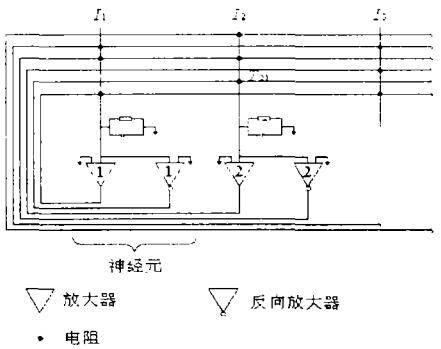

图 3　Hopfield 模拟电路

$f_i(\bullet)$ 是 S 形函数，故 $f_i(\bullet)$ 单调增，上式右边的每一项都是非负的，从而

$$\frac{dE}{dt} \le 0 \quad (11)$$

并且，仅当[2-3] $dv_i/dt = 0, \forall i$ 时，等号成立。dv_i/dt 对应的是状态空间中能量函数 E 的稳定平衡点，表示的是网络最终可能的输出值的集合。因为函数 E 是有界函数，故上式表明网络总是吸引到 E 函数的局部最小值上。

通过适当地选取权 w_{ij} 的值以及外部输入信号 I_i，将优化问题匹配到神经网络上。神经网络在进行这样的构造后，给输入电压一组初始值，这时，网络将收敛到极小化目标函数 E 的稳定状态，目标函数达到它的局部极小值。

3) 自组织网络

它以 ART 模型、Koholon 模型为代表，用于聚类。人工神经网络具有分布式存储信息、并行地处理信息

和进行推理、以及自组织自学习等特点，解决了众多用以往方法很难解决的问题。

径向基函数神经网络 RBF 在克服局部极值及提高拟合精度方面具有良好的效果，得到广泛的应用。

3.2 决策树[3]

1）鉴别生成侯选子树：使用一个调整的错误率，AE(T) = E(T) + aleaf_count(T)。

2）对子树的评估：通过 test set 找到最佳子树。

3）最佳子树进行评估：使用 evaluation set。

4）考虑代价(cost)的问题。

2.4 进化式程序设计(Evolutionary programming)[19,24-25]

这种方法的独特思路是，系统自动生成有关目标变量对其他多种变量依赖关系的各种假设，并形成以内部编程语言表示的程序。内部程序(假设)的产生过程是进化式的，类似于遗传算法过程。当系统找到较好地描述依赖关系的一个假设时，就对这程序进行各种不同的微小修正，生成子程序组，再在其中选择能更好地改进预测精度的子程序，如此依次进行，最后获得达到所需精度的最好程序时，由系统的专有模块将所找到的依赖关系由内部语言形式转换成易于为人们理解的显式形式，如数学公式，预测表等。

2.5 基于事例的推理方法(Case based reasoning CBR)[14,19,26]

这种方法的思路非常简单，当预测未来情况或进行正确决策时，系统寻找与现有情况相类似的事例，并选择最佳的相同的解决方案，这种方法能用于很多问题求解，并获得好的结果，其缺点是系统不能生成汇总过去经验的模块或规则。

2.6 遗传算法(Genetic Algorithms: GA)[3,12,22]

用于分类，关联规则采掘等。遗传算法模仿人工选择培育良种的思想，从一个初始规则集合(知识基因)开始，逐代地通过交换对象成员(杂交、基因突变)，产生群体(繁殖)，评价并择优复制(适者生存，不适应者淘汰)，逐代积累计算，最终得到优化的知识集。但是基本遗传算法局部搜索能力差，存在早熟收敛现象，而且导致算法的收敛性能差，特别是后期搜索迟钝，不能保证达到局部最优，因此董聪提出广义遗传算法。

2.7 非线性回归方法[22,27-28]

一个非线性回归模型中参数的出现是非线性的，例如：

$$Y_t = X_t^{\theta} + \varepsilon_t \qquad (12)$$

其中参数 θ 是要估计的。与线性模型类似，可用最小二乘法极小化(12)式

$$S(\theta) = \sum (Y_t - X_t^{\theta})^2 \qquad (13)$$

来估计 θ，用 S 代替 $S(\theta)$ 以简化记号，S 的最小值可由上式对 θ 求微分得到，让导数等于零，如下面：

$$\frac{\partial S}{\partial \theta} = -2\sum (Y_t - X_t^{\theta})^2 (\log X_t) X_t^{\theta} = 0 \qquad (14)$$

然后设法求出 θ 的解，其解用 $\hat{\theta}$ 表示。然而它不能求出的明显的表示式。作为代替，整理后的结果为方程：

$$\sum Y_t (\log X_t) X_t^{\theta} = \sum (\log X_t) X_t^{2\theta} \qquad (15)$$

从某一假定的 $\hat{\theta}$ 值开始仅用迭代的方法就可以产生 LS 估计 $\hat{\theta}$[19]。

这种方法的基础是，在预定的函数的基础上，寻找目标度量对其他多种变量的依赖关系，在金融市场或医疗诊断的应用场合，比较好的提供可信赖的结果。

2.8 粗糙集理论[29]

1990 年波兰华沙理工大学 Z. Pawlak 教授提出了粗糙集(Rough Sets)，它是基于集合理论，对人们获得的大量现实数据进行分类，从中发现隐藏在数据中各种信息，为解决不完整、不精确和不确定性问题的数据挖掘问题提供了一个科学的方法，近年来成为国内外研究的一个热点。

3 数据挖掘技术应用及研究展望

数据挖掘作为一门新型的交叉学科，有着广泛的应用前景，无论是金融界、医疗界、保险业还是股票市场和 Internet 领域，到处都有数据挖掘的足迹。近几年来随着网络的普及和电子商务的日益成熟，数据挖掘的应用越来越广泛。就目前的应用看，数据挖掘算法和技术可概括地分为下面几种使用类型：[19,28,31]

1）关联发现[31]

关联的一个典型例子是市场菜篮子分析，此分析与一组产品相关联。通过挖掘事物数据可发现关联规则，利用此规则可以了解客户的行为。例如，观察客户对办公用品的订货，在那些订购笔的客户中，有 70% 也订购了写字台，“笔”是规则的前提，“写字台”是规则的结果，关联规则中可有任意多个前提和结果。挖掘系统试图在给定的数据集中找到尽可能多的关联规则或模式。此外，“70%”表示了置信度因子。分析员通常感兴趣的是一组关联，如：查找所有以“写字台”作为结果的关联，以便制定战略来增加“写字台”的销售，因为“写字台”是一种高利润的商品。如果“铅笔”是一种微利商品，那么要检查所有以“铅笔”作为前提的关联，以便确定不再销售铅笔时带来的影响。如果“写字台”具有很高的利润，那么需要查找所有以“铅笔”作为前提，并以“写字台”作为结果的关联，以便了解为增加“写字台”订货所需的“铅笔”。事务分析不一定要同时处理所有订货，只要在给定时间框架内包

含所有订货就可以了[20,32]。

2）聚类分析

当要分析的数据缺乏描述信息，或者是无法组织成任何分类模式时，利用聚类算法可以自动地找到类。聚类功能可用于一组顾客的现金流分析，这些顾客在一月的特定时间内付帐（例如，当收到社会保险支票时，或者月工资存入帐户时），聚类还可用于市场细分，寻找相关的组。[3,25]。

3）分类

分类问题涉及规则的查找，此规则将数据记录划分成不连贯的组，划分基于数据记录的属性。例如：信息认可和商店定位。在商店定位中，首先按成功的商店、一般商店和失败商店进行排列，然后得出这三类商店各自具有特殊性，然后选择包含位置属性的地理数据库，并对每一项预期的商店位置属性进行分析，以便确定预期的商店定位属于哪一类。只有那些符合成功一类要求的商店才选作所希望的商店定位。

4）神经网络的使用

神经网络已应用于许多商业领域，例如：市场营销——此领域需要检查客户的行为以便构造微观市场细分和邮寄表，并且还要寻找理想的客户群。财经分析——此领域包括现金流分析和欺诈检查。商业运作——此领域包括传送计划和后勤分析。[22,23]。

5）规则发现和决策树

使用规则发现算法用于带有属性或描述的数据项中，其目的是要显式描述抽取的规则。显式规则分析员必须明确的理解并指明规则，它需要指明的是“好的”和“差的”信用风险客户。没有恰当的解释就“拒绝信用请求”会带来很高的风险。[3,15,26,31]

6）顺序模式和顺序序列

在许多情况下，客户事务要经历很长时间，这也是顾客全局关系的一部分。分析员对订货之前发生的事情感兴趣，例如，邮寄宣传材料、订购本身、客户服务请求、定单传送（及时或非及时），售后服务、后继订货以及其他与顾客交互。数据库可以包含所有这些临时数据，并可跨跃多个时间段。顺序模式功能可以分析数据库中一组此类型的数据，并发现某一段时间内顾客购买定单中常见的模式。例如，顾客现在订购了一台打印机，在以后还可能定购“打印纸”。定货应具有一种基于打印机生命周期的模式：初始购买、售后服务和维修服务。可能在初始购买时有大量订货，在服务请求时是限量的，而在每一服务请求完后又有大量定货。在售后管理和收入计划中，关于这些模式的知识是有价值的。此外，分析员可发动促销活动，以便将这种模式改变或更符合效益要求的模式，并增加客户的满意度。顺序模式和顺序序列可看成一种特定的关联规则。此关联模式用于查找一组客户，这些客户符合特定频率的购买模式。[1,6,16,27]

数据挖掘技术除了以上的一些重要的应用之外，还存在一些问题[3]。

1）数据质量　由于是数据驱动，而且相对于不接受管理，因此很容易遇到数据质量的问题。许多数据库很可能是动态的、有错误而且不完整的、有冗余、稀疏的、当然也是很大的。因此在恰当使用知识发现功能和技术的同时，还要小心的分析异常。

2）数据可视化　将数据库大量的数据可视化需要复杂的数据可视化工具。它有助于分析员增加人们的视觉能力，尤其是数据维数较低的时候。由于数据库中的数据量非常巨大，很容易使分析员变得不知所措。数据挖掘可通过设定富有成效的探索的始点并按恰当的隐喻来表示数据给予帮助。[3,33]。

3）极大数据库的问题　数据需要事务数据和细节数据，以便了解顾客的行为和购买模式。极大数据库除了在进行系统管理时存在问题以外，许多挖掘系统也由于极大的数据库尺寸而存在问题。查询数据的尺寸很可能对特定技术（例如神经网络训练）造成困难。在许多情况下，需要使用其它的数据抽样技术。

4）性能和成本　为了满足许多数据挖掘系统的计算要求，需要在硬件、操作系统软件和数据库系统采用并行技术。这些资源大大增加了成本，并且使并行型技术专家构成的信息技术资源也变得紧张。

5）信息分析员的技能　信息分析需要丰富的领域知识，并具有很强的调查能力，同时还应用创造性。创造性允许分析员试验各种知识发现技术，以便发现大量潜在的模式和关系，然后分析并了解它，最后生成预测模型，并按易理解的形式发布。

展望数据挖掘的出现只有短短的几年时间，如今方兴未艾。数据挖掘技术所表现出的广阔应用前景吸引了众多的研究人员和商业公司。一批数据挖掘系统被开发出来，并在商业、经济、金融、管理等领域都取得了应用性成果。采用的方法综合了机器学习、模式识别、统计学、知识发现、数据库和数据分析等领域的研究成果。但总的说来，这些系统基本上还停留在实验阶段，在适应性、系统效率方面还不尽人意。随着硬件环境、挖掘算法的改进及应用经验的积累，数据挖掘技术与应用将会得到长足的进展。

参考文献：

[1]　吴少敏，马建生，陈贻龙，等. 实用数据挖掘系统[J]. 冶

金自动化. 2002,(1):6 - 10.
[2] 陈莉,焦李成. Internetweb 数据控制研究形状及最新进度[J]. 西安电子科技大学学报. 2001,28(1):114 - 119.
[3] 王燕,李睿,李明. 数据挖掘技术应用研究[J]. 甘肃科技,2001,17(1):49 - 50.
[4] 丁夷. 数据挖掘——技术与应用综述[J]. 西安邮电学院学报. 1999,4(3): 41 - 44.
[5] 陈莉, 焦李成. Internet/Web 数据挖掘现状及最新进展[J]. 西安电子科技大学学报. 2001,28(1):114 - 119.
[6] DASKALAKI S, KOPANAS I, GOUDARA M, etal. Data Mining for Decision Support on Customer Insolvency in Telecommunications Business. [J]. European Journal of Operational Research, 2003, 145(2): 239 - 255.
[7] ANANTHANARAYANA V S, NARASIMHA MURTY M, SUBRAMANIAN D K. Tree Structure for Efficient Data Mining Using Rough Sets[J]. Pattern Recognition Letters. 2003, 24(6):851 - 862.
[8] ZHONG NING, DONG JU-ZHEN, SETSUO OHSUGA. Meningitis Data Mining by Cooperatively Using GDT - RS and RSBR [J]. Pattern Recognition Letters. 2003, 24 (6):887 - 894.
[9] KAO S C ,CHANG H C , LIN C H. Decision Support for the Academic Library Acquisition Budget Allocation via Circulation Database Mining [J]. Information Processing & Management, 2003, 39(6):133 - 147.
[10] ZHOU ZHI-HUA . Three Perspectives of Data Mining[J]. Artificial Intelligence. 2003, 143(1):139 - 146.
[11] Hu Yi-Chung , Chen Ruey-Shun , Tzeng Gwo - Hshiung Finding Fuzzy Classification Rules Using Data Mining Techniques[J]. Pattern Recognition Letters, 2003, 24(1 - 3):509 - 519.
[12] 王珊. 数据仓库技术与联机分析处理[M]. 北京:科学出版社. 1999.
[13] ARHTUR L HSU, SAMAN K HALGAMUGE. Enhancement of Topology Preservation and Hierarchical Dynamic Self - organising Maps for Data Visualization[J]. International Journal of Approximate Reasoning, 2003, 32: 259 - 279.
[14] 王小平,曹立明. 遗传算法——理论应用与软件实现[M]. 西安交通大学出版社,2002.
[15] ZHANG SHICHAO, ZHANG CHENGQI, YAN XIAOWEI. Post - mining: Maintenance of Association Rules by Weighting[J]. Information Systems,2003,28(7):691 - 707.
[16] VLADAN BABOVIC, JEAN - PHILIPPE DRECOURT, MAARTEN KEIJZER, et al. Adata Mining Approach to Modelling of Water Supply Assets [J]. Urban Water. 2002,4(4):401 - 414.
[17] CHEN GUOQING ,WEI QIANG . Fuzzy Association Rules and Extended Mining Algorithms [J]. Information Sciences, 2002, 147(1 - 4):201 - 228.
[18] FERNANDO ALONSO, JUAN P. CARACA - VALENTE, AAGEL L. GONZALEZ , et al. Combining Expert Knowledge and Data Mining in a Medical Diagnosis Domain[J]. Expert Systems with Applications, 2002, 23 (4): 367 - 375.
[19] CHRIS RYGIELSKI, WANG JYUN-CHENG, DAVID C. YEN. Data Mining Techniques for Customer Relationship Management[J]. Technology in Society, 2002, 24(4): 483 - 502.
[20] HONG TZUNG-PEI, LIN KUEI-YING , WANG SHYUE-LIANG. Fuzzy Data Mining for Interesting Generalized Association Rules[J]. Fuzzy Sets and Systems, 2003, 138 (2):255 - 269.
[21] LIAN J, LAI X M, LIN Z Q, et al. Application of Data Mining and Process Knowledge Discovery in Sheet Metal Assembly Dimensional Variation Diagnosis[J]. Journal of material processing Technology, 2002, 129 (1 - 3): 315 - 320.
[22] 徐铭杰. 空间数据挖掘模型和方法研究[J]. 河南纺织高等专科学校学报, 2002,14(1):15 - 17.
[23] ALEX BERSON, STEPHEN SMITH, KURT THEARLLING. 构建面向 CRM 的数据挖掘应用[M]. 贺奇,郑岩,魏黎,等译. 北京:人民邮电出版社. 2001.
[24] 陈捷,唐世渭,杨冬青,等. 面向移动环境的时空数据库研究现状与展望[J]. 计算机工程与应用,2002,38 (16):1 - 3
[25] 杨会志. 数据挖掘技术的主要方法及其发展方向[J]. 河北科技大学学报, 2000,21(3):77 - 80.
[26] HAN JIAWEI, MICHELINE KAMBER. Data Mining: Concepts and Techniques[M]. San Francisco: Morgan Kaufmann Publishers. 2000.
[27] 吴载斌,王斌会. 数据挖掘软件的介绍及其评价[J]. 计算机时代, 2002,(7):3 - 4.
[28] 刘力扬. 数据挖掘与数据库知识发现[J]. 河南电大学报, 2000,(3):42 - 43.
[29] 仲红. 数据挖掘技术的深入研究[J]. 淮南工业学院学报. 2002,22(2):42 - 45.
[30] PAWLAK Z, ROUGH SETS. Theoretical Aspects of Reasoning about Data[M], Dordrecht, boston, Londom: Kluwer Academic Publisbers,1991.
[31] 陈岚岚,杨波,李旭霞. 数据挖掘技术及其发展方向[J]. 武警工程学院学报,2002,18(4):13 - 15.
[32] RAKOWSKY DA 著. 非线性回归模型统一的实用方法[M]洪再吉,韦博成,吴诚欧,等译. 南京:南京大学出版社. 1986.
[33] 郑君里,杨行峻. 人工神经网络[M]. 北京:高等教育出版社,1992.

Development in Modern Data-Mining Techniques

LIANG Xie-xiong, LEI Ru-huan, CAO Chang-xiu

(College of Automation, Chongqing University, Chongqing 400030, china)

Abstract: Data-mining is a composite and multi – disciplinary technology newly developed. This paper will start by describing the fundamentals and basic methods of the technology followed by an overview of the latest development of its applications, problems and hurdles currently facing and the way forward and future work.

Key Words: knowledge discovery in database, data-mining, data-warehouse, decision support

(编辑　吕赛英)

(上接第 20 页)

Elastic Buckling and Critical Load of an Pre – stressing Arch

ZHANG Pei-yuan, SHENG Tian-wen, ZHANG Xiao-min

(College of Resources and Environmental Science, Chongqing University, Chongqing 400030, china)

Abstract: According to the field theory of additional deformation on pre – stressed configuration, in the paper, the ordinary expression of the governing equation and variational equation of elastic buckling are brought forward。 Under the theory system, through lowering dimensions, the governing equation and variational equation for the critical condition solution of elastic buckling of a plane arch are deduced, and the eigenvalues problem of the linear homogeneous differential equations corresponding to the equations are concluded While Abandoning the plane assumption and considering shearing deformation, the linear finite element method arithmetic of bending bar's cross section containing six degree of freedom is given. The process of derivation and calculational results show that, under this system info, the finite element equations of bending bar deduced are accurate and easy to be used to numeric calculations, and the conclusion achieved is more practical.

Key words: pre-stressed configuration; field theory of additional deformation; elastic buckling; arch; shearing deformation

(编辑　成孝义)