**数据挖掘的研究与应用进展**

---应宇杰19151633

## 引 言

在社会与经济生活中,任何组织经过若干年的发展 ,都会积累海量的、以各种形式存储的数据 ,但要从中发现有价值的信息或知识 ,改变“数据丰富 ,信息缺乏”尴尬境地 ,却是非常艰巨的. 为弥合数据与信息之间的鸿沟 ,数据挖掘(Data Mining) 应运而生 ,使人们能从海量数据中发现感兴趣的知识 ,而这些知识是隐含的、事先未知的信息.

## 数据挖掘与知识发现

**1**. **1 知识发现与数据挖掘的兴起**

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. 在传统的决策支持系统中 , 知识库中的知识和规则是由专家和程序人员建立的 ,是由外部输入的. 对于那些决策者明了的信息 ,可以用查询、联机分析处理直接获取 ,而另外一些隐藏在大量数据中的关系、趋势 ,是既难以发现又至关重要的 ,这类问题就可以用数据挖掘来处理. 知识发现的研究经历了从ML(机器学习)到MD(机器发现) 到KDD(Knowledge Discovery inDatabase 基于数据库的知识发现)几个阶段.KDD 主要研究①定性知识和定量知识的发现;②数据汇总③知识发现方法④数据依赖关系的发现和分析⑤发现过程中知识的应用 ;

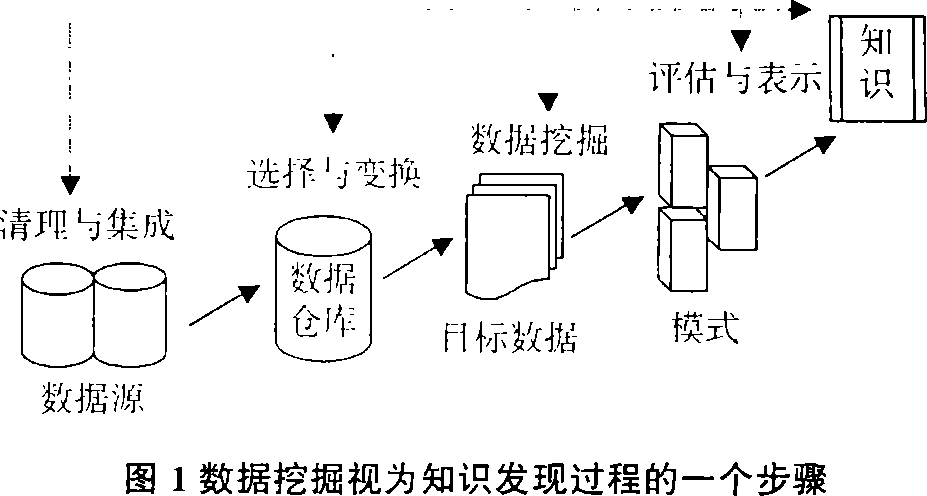
⑥集成的交互式的知识发现系统⑦知识发现的应用.

**1**. **2 数据挖掘的定义**

数据挖掘(Data Mining),又译作数据采集 ,一种公认的定义是WJFrawley,G. PiatetskyShapiro 等人提出的:数据挖掘 ,就是从大型数据库的数据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 ,提取的知识可表示为概念(Concepts) 、规则( Rules) 、规律( Regularities) 、模式( Patterns) 等形式. 此定义把数据挖掘的对象仅定义为数据库. 广义地讲:数据挖掘是在一些事实或观察的集合中寻找模式的决策支持过程. 也就是说 ,数据挖掘的对象不仅是数据库 ,也可以是文件系统 ,或其它任何数据集合.

从数据挖掘的定义可以看出 ,数据挖掘和数据库知识发现 KDD 具有很大的重合度 ,甚至许多学者认为数据挖掘和数据库知识发现是等价的概念 ,人工智能(AI) 领域习惯称 KDD ,而数据库领域习惯称数据挖掘.

长期以来 ,在知识发现领域这两个术语的范畴和使用界限一直不很清晰.直到 KDD96 国际会议上 ,知识发现研究领域的知名学者 Fayyad,Piatetsky- Shapiro和Smyth 就这两个术语的关系作了如下阐述: KDD 是指从数据库中发现知识的全部过程,即识别出存在于数据库中有效的、新颖的、具有潜在效用的乃至最

终可理解的模式的非平凡的过程. Data Ming 则是全部过程中的一个特定步骤. 这样 ,知识发现的过程可描述为: ①数据清理 : 消除噪声或不一致数据; ②数据集

成:把多种数据源组合在一起; ③数据选择 : 从数据库检索与分析任务相关的数据; ④数据变换 :数据变换或统一成适合挖掘的形式;⑤数据挖掘 :使用智能方法发现或提取数据模式; ⑥模式评估 : 根据某种兴趣度度量,识别表示知识的真正有趣的模式;⑦知识表示:使用可视化等知识表示技术 ,向用户提供挖掘到的知识. 可表示为图1.

**1.3数据挖掘的特点**

* 1. 挖掘对象是海量的、复杂的各种类型的数据. 这些源数据可能是残缺、充斥噪音的“脏”数据 ;



* 1. 挖掘的结果是潜在的、未知的、多样性的(发现的知识可以是多种形式的) ;
  2. 挖掘方法是不确定的. 数据挖掘方面没有所谓最好的技术或通用的技术 ,因此 ,问题不是某一种方法比另一种方法更好 ,而是哪一种更适合所要解决的问题;
  3. 数据挖掘支持在线数据存取;
  4. 技术的综合性. 数据挖掘融入了人工智能技术、数据库技术、数理统计技术、可视化技术等技术和哲学、逻辑学等学科的知识.

**1**.**4数据挖掘的任务及挖掘对象**

数据挖掘的任务是从大量的数据中发现知识,数据挖掘是知识发现的核心技术数据挖掘的主要的目标是发展有关的方法论、理论和工具 ,以支持从大量的数据中提取有用的、让人感兴趣的知识. 知识最主要的表示形式是模式( Patterns) . 模式是用语言 *L* 来表示的一个表达式 *E ,* 它可用来描述数据集 *F* 中数据的特性 *, E* 所描述的数据是集合 *F* 的一个子集*FE.E* 作为一个模式要求它比所列举数据子集 *FE* 中所有元素的描述方法简单. 模式有很多种 ,按功能可分为两大类:预测型( Predictive) 模式和描述型(Descriptive) 模式.

数据挖掘功能以及它们可以发现的模式类型主要有以下几种: ①总结规则( Summarization rule) ; ②关联规则(Association rule) ; ③分类规则(Classification rule) ; ④聚类规则(Clustering rule) ; ⑤预测分析( Predic2 tion analysis) ; ⑥趋势分析( Trend analysis) ; ⑦偏差分析(Deviation analysis) .

原则上讲 ,数据挖掘可以在任何类型的信息存储设备上进行 ,它包括: ①关系数据库 ; ②数据仓库(Da2 ta Warehouse) ; ③事务数据库( Transactional Database) ; ④高级数据库系统 :包括面向对象数据库;对象 —关系数据库;空间数据库;时间数据库;时间序列数据库;文本数据库;多媒体数据库;异种数据库;Web 数据库.

**1**.5**数据挖掘的算法和技术**

**1.5.1数据挖掘的算法**

数据挖掘算法是对数据挖掘算法的具体实现,一般由三部分组成(下文中的模型上指从数据库中发现的模型) :

1、模型表示. 用于描述要发现的模型是语言. 如果语言的描述能力较强 ,就有助于发现精确的数学模型. 常用的模型表示方法有决策树、非线性回归、基于事理的推理、贝叶斯网络和归纳程序设计等方法.

2、模型评价标准. 对一个所发现的模型在多大程度上符合发现目的要求做出定量的评价. 对预测类的模型 ,可以利用一些测试数据集来评价其精确度. 对描述类的模型 ,可以在精确度、新颖性、实用性及可理解性等多个方面进行评价.

发现方法. 分为参量发现和模型发现. 在模型表示和模型评价标准被确定之后 ,数据挖掘就完全变成了一个优化任务 ,即从数据的描述中发现最适合评价标准的参量或模型. 具体而言 ,参量发现就是在确定数据集和模型表示之后 ,寻找最适合模型评价标准的参量. 模型发现是一个循环地试探过程 ,需要不断更改模型表示 ,最后确定出恰当数量的模型一般来说 ,不存在一个普遍适用的算法 ,在实际应用中 ,要针对特定的领域 ,选择有效的数据挖掘算法.

**1.5.2数据挖掘的常用技术**

人工神经网络:它从结构上模仿生物神经元结构 ,是一种通过训练来学习的非线性预测模型 ,可以完成分类、聚类、特征采掘等多种挖掘任务. 其主要缺点是知识的表示不够清晰 ,用加权链接单元的网络表示的知识很难让人理解. 因此 ,当前神经网络算法的一个研究热点就是提取隐藏在训练神经网络中的知识 , 并象征地解释这些知识 ,已有的算法是网络提取算法和灵敏度分析.

遗传算法:这是模拟生物进化过程的算法 ,由三个基本算子组成: ①繁殖 (选择) 是从一个旧种群(父代) 选出生命力强的个体 ,产生新种群(后代) 的过程; ②交叉(重组) 选择两个不同个体(染色体) 的部分(基因) 进行交换 ,形成新个体; ③变异(突变) 对某些个体的某些基因进行变异(1 变 0 、0 变 1) . 在数据挖掘中 , 它还可以用于评估其它算法的适合度.

最近邻技术:这种技术通过 *K* 个最与之相近的历史记录的组合来辨别新的记录 *,* 有时也称这种技术为 *K* —最近邻方法. 最近邻分类是基于要求的或懒散的学习法 ,即它存放所有的训练样本 ,并且直到新的



(未标记的) 分类需要时才建立分类.

规则归纳:规则反映数据项中的某些属性或数据集中某些数据项之间的统计关系. 比较典型的规则算法是AQ 系列算法、洪家荣改进的 AQ15 方法 ,以及洪家荣的AE5 方法. 近年来研究得更多的是关联规则算法. 关联规则发现任务是指:给定一个事务数据库 *D ,* 求出所有满足最小支持度和最小可信度的关联规则*.* 该问题可分解为两个子问题 *:* ①求出 *D* 中满足最小支持度的所有常用物品集; ②利用常用物品集生成满足最小可信度的所有关联规则. 其关键技术是第一步. Apriori 算法是一种有效的关联规则挖掘算法. 频繁模式增长( FP - 增长) 是一种不产生侯选的挖掘频繁项集方法. 现在的研究热点是多维关联规则的挖掘和基于约束的关联规则挖掘.

贝叶斯分类:是基于贝叶斯定理的统计学分类方法 ,具有高准确率与高速度等特点. 贝叶斯分类常用的方法有两种:朴素贝叶斯分类和贝叶斯信念网络. 前者假定一个属性值对给定类的影响独立于其它属性的值(既类条件独立) ,后者说明联合条件中的概率分布 ,它提供一种有向的非循环图 ,结点表示属性变量 , 边表示属性之间的概率依存关系 ,与每个结点相关的是条件概率分布 ,它描述该结点与父结点之间的关系.

粗集(Rough Set) 方法:粗集方法用于处理离散值属性的含糊性和不确定性. 粗集是由集合的下近似、

上近似来定义的. 下近似中的每个成员都是该集合的确定成员 ,而不是上近似中的成员肯定不是该集合的成员 ,可以认为粗集是具有三值隶属函数的模糊集 ,即是、不是、也许.

模糊集方法: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. 模糊集不仅可以处理不完全数据、噪音或不精确数据 ,而且在开发数据的不确定性模型方面是有用的 ,与传统的方法相比可提供更灵巧、更平滑的性能. 一般地 ,模糊逻辑在基于规则的系统中的使用涉及:将属性值转换成模糊值、对于给定的新样本可以使用多个模糊规则、组合上面的和得到一个系统返回的值.

概念树方法:数据库中记录的属性字段按归类方式进行抽象 ,建立起来的层次结构称为概念树. 利用概念树提升的方法可以大大浓缩数据库中的记录. 对多个属性字段的概念树提升 ,将得到高度概括的知识基表 ,然后再将它转换成规则.

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 , 求得相应的数学公式. 比较典型的BACON 发现系统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.

另外还有统计分析方法、可视化技术和决策树法等.

## 2.数据挖掘的应用领域及商业产品介绍

数据挖掘技术的潜在应用领域非常广泛,从政府决策、科学研究到商业经营决策都可以采用这一新兴技术. 下面仅列举了一些当前比较活跃的数据挖掘应用方向:

**2.1商业领域的应用**

* 1. 零售业

数据挖掘在零售业的应用分为两类:数据库行销(Database marketing) 和货篮分析(Basket analysis) ,前者的任务是通过交互式查询、数据分割和模型预测来选择潜在的顾客 ,以便有的放矢地向他们推销产品; 后者的任务是分析市场销售数据(如 POP 数据库) 以识别顾客的购买行为模式 ,从而有助于确定商店货架的布局排放以促销某些商品 ,同时对进货的选择和搭配上也更具目的性.

这方面的系统有:CoverStory 系统 ,美国信息资源公司与MIT Sloan 管理学院的Little 合作开发的一个商业系统 ,该系统可以标识并排列产品量跨时间、跨地区的富有意义的变化、并对这些变化提供一些可能的解释 ,最终以自然语言的形式输出. Opportunity Explorer 系统 ,可用于超市商品销售异常情况的因果分析.

* 1. 金融业



数据挖掘在金融业的应用分为两类:金融投资和欺诈甄别. 典型的金融投资分析领域有投资评估和股票交易市场预测 ,分析方法一般采用模型预测法(如神经网络或统计回归技术) . 这方面的系统有 Fidelity Stock Selector ,LBS Capital Management . 前者的任务是使用神经网络模型选择投资 ,后者使用了专家系统、神经网络和基因算法技术辅助管理有价证券. 针对银行或商业上经常发生的欺诈行为 ,如恶性透支等而开发的成功系统有: FALCON 系统和 FAIS 系统 , FALCON 系统是 HNC 公司开发的信用卡欺诈估测系统. FAIS 系统是一个用于识别与洗钱有关的金融交易系统 ,它使用的是一般的政府数据表单.

**2.2科学领域**

美国加州理工学院(CTT) 喷气推进实验室与天文科学家合作开发了一个天体图像分类分析工具 SKI2 CAT(sky image cataloging and analysis tool) ,使用 SKICAT 对天体数据进行分析 ,一方面是通过机器学习将知识提取过程由学习算法完成 ,从而实现对大批量数据的分析;另一方面是辩识那些亮度很低、人工难以判读的天体图像 ,以进行后续分析.

Simon 等人研制的BACON 系统成功地重新发现了理想气体定律、库伦定律、开普勒第三定律、欧姆定律和伽利略定律.

同时 ,还有诸如中国科学院计算机技术研究所与国家税务部门合作 ,开发了计算机选案系统 ,用于稽查和追缴偷、漏、欠税款 ,查处和纠正纳税人的违法行为. 由Acknosoft 公司用 KATE 发现工具开发的 CAS2

SIOPEE 系统 ,已用于诊断可预测在制造波音飞机过程中可能出现的问题.

在通信网络管理方面 ,芬兰 Helsinki 大学与一家远程通信设备制造厂合作的 TASA 系统 ,可用于网络故障的定位检测和严重故障的预测等任务中.

## 2.3数据挖掘面临的挑战及发展趋势

当前,数据挖掘面临的主要挑战是: ①数据输入形式的多样性 ; ②数据挖掘算法的有效性与可测性 ; ③ 用户参与和领域知识; ④证实技术的局限 ; ⑤知识的表达和解释机制 ; ⑥知识的维护和更新 ; ⑦私有性和安全性支持的局限、与其它系统的集成.

所以 ,当前数据挖掘的研究是以知识发现的任务描述、知识评价与知识表示为主线 ,有效的知识发现算法为中心 ,面向具体应用 ,开发原型系统与实用系统 ,研究与开发基于数据挖掘的通用工具.

因此 ,数据挖掘领域的发展趋势可概括为:

1. 应用领域的探索和扩张. 在注重理论、技术研究的同时 ,强调实际应用研究 ,如一般化的、通用的及针对特定领域的数据挖掘系统的开发.

算法的效率和可伸缩性. 数据挖掘直接面对的是海量数据 ,且这些数据之间已含着各种繁杂关系 ,这就导致挖掘过程中搜索空间和搜索维数的激增 ,且其间的许多不确定因素和干扰因素也就随之增加 ,但许多成熟的算法是基于内存的 ,这就对算法的效率提出了严峻的挑战. 另一方面 ,由于数据量是随时间激增的 ,因此 ,针对单独、集成的数据 ,挖掘功能的可伸缩性就显得非常必要.

1. 数据挖掘系统的交互性. 数据挖掘中适当的用户参与是必不可少的 ,是基于以下几个方面的原因: ①友好的、完善的交互界面是用户准确表达其要求和挖掘策略的保证 ; ②用户的背景知识和指导作用可以提高挖掘效率 ,并保证发现知识的有效性. 目前一个重要的研究方向就是所谓基于约束的挖掘(con2 straint - bad) 它致力于在增加用户交互的同时如何改进挖掘处理的总体效率; ③通过交互界面 ,系统可以直接、有效地获取用户的感兴趣模式 ,从而提高挖掘的有效性.
2. 复杂数据源和数据类型的处理. 数据挖掘系统的理想体系结构是与数据库和数据仓库系统的紧耦合方式 ,把事务管理、查询处理、联机分析处理和联机分析挖掘集成在一个统一的框架中 ,然而 ,正如前所述 ,实际应用中的数据挖掘对象是各种类型的数据库 ,甚至是没有完整数据结构的数据集 ,因此 ,如何把这些特殊数据类型的专用分析方法与现在成熟的基于关系数据库和数据仓库的数据挖掘方法集成起来 , 实现这些复杂数据源和数据类型的处理 ,是一个重要的发展方向.



1. 隐私保护与信息安全. 在发展数据挖掘的同时 ,需要进一步开发有关方法以便在适当的信息访问和挖掘过程中保护隐私和信息安全.
2. Web 挖掘. 一般地Web 挖掘可分为三类:Web 内容挖掘(Web content mining) 、Web 结构挖掘(Web structure mining) 和 Web 使用记录挖掘(Web usage mining) . 目前 Web 挖掘研究的主要方向包括:Web 文本的自动分类、多层次 Web 信息库的建立、Web log 挖掘 ,以及其它涉及信息安全、搜索的效率、查询结果的质量、搜索工具的可伸缩性等方面的问题研究. 可以预见 ,随着 XML 的兴起 ,Web 页面会蕴涵更多的结构化和语义信息 ,这会使 Web 挖掘变得更有效.

## 结论

综上所述,数据挖掘的发展 ,为人工智能领域中解决知识获取这一难题提供了有效的解决方案 ,同时也为数据库领域中对数据的深层次利用开辟了新的发展空间. 然而就现状而言 ,数据挖掘仍有一定的局限性 ,如系统的低性能和挖掘对象的单一性等 ,但可预见 ,随着研究的进一步深入和数据存储及表达方式的日趋标准化 ,数据挖掘将变得更为有效并将得到更为广泛的应用