可拓学丛书

可拓集与可拓数据挖掘

蔡　文　杨春燕
陈文伟　李兴森　著

科学出版社
www.sciencep.com

可拓学丛书

可拓集与可拓数据挖掘

蔡　文　杨春燕　陈文伟　李兴森　著

国家自然科学基金资助项目

广东省自然科学基金资助项目

科学出版社

北京

内 容 简 介

可拓数据挖掘以可拓集为集合论基础，结合可拓方法与现有数据挖掘方法去挖掘数据库或数据仓库中基于可拓变换的知识，为经济、金融、管理、营销、策划、医学、设计等领域的决策和技术创新提供依据. 本书是第一本可拓数据挖掘的专著，提出了研究这一领域的理论基础、方法体系和应用范围，并给出简单、浅显的实用案例. 本书理论与应用相结合，分析透彻. 为方便不同知识背景和不同层次读者的学习，书中配备了通俗易懂的案例.

本书适合高等院校师生、工程技术人员和管理决策人员阅读，特别适合作为高等院校相关专业本科、硕士、博士生的选修课教材.

图书在版编目(CIP)数据

可拓集与可拓数据挖掘/蔡文等著. —北京：科学出版社，2008
(可拓学丛书)
ISBN 978-7-03-021817-9
Ⅰ.可… Ⅱ.蔡… Ⅲ.拓扑-研究 Ⅳ.O189
中国版本图书馆 CIP 数据核字(2008) 第 061252 号

责任编辑：范庆奎 房 阳／责任校对：刘小梅
责任印制：赵德静／封面设计：王 浩

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com
双青印刷厂 印刷
科学出版社发行 各地新华书店经销
*
2008 年 6 月第 一 版 开本：A5(890 × 1240)
2008 年 6 月第一次印刷 印张：5 1/2
印数：1—2 500 字数：199 000
定价：30.00 元
(如有印装质量问题，我社负责调换〈双青〉)

《可拓学丛书》序

人类的历史, 是一部解决矛盾问题、不断开拓的历史. 可拓学研究用形式化的模型分析事物拓展的可能性和开拓创新的规律, 形成解决矛盾问题的方法, 对于提高人类智能有重要的意义. 根据这些研究成果探讨用计算机处理矛盾问题的理论和方法对于提高机器智能的水平有重要的价值. 可拓学研究正是基于这种目的而进行的.

可拓学选题于 1976 年, 1983 年发表首篇论文 "可拓集合和不相容问题". 十多年来, 经历了无数的艰辛, 在广大可拓学研究者的努力下, 逐步形成了可拓论的框架, 开展了在多个领域的研究, 一个新学科的轮廓已经形成.

近年来, 不少学者加入了建设这一新学科的行列. 可拓学的应用研究和普及推广迫切需要一批介绍可拓学的书籍, 供研究者参考. 为此, 我们组织了《可拓学丛书》的编写, 希望通过这套丛书, 把可拓学介绍给广大学者.

诚然, 目前可拓学还未完全成熟, 可拓学的研究水平还不高, 理论体系还要进一步建设, 应用研究还需深入进行, 大量的问题尚待解决. 因此, 这套丛书只能起抛砖引玉的作用. 我们希望通过这套丛书, 为广大学者提供可拓学的初步知识和可拓学的思维方法, 并提供研究的课题.

我们相信, 丛书的出版将会吸引更多学者加入可拓学的研究行列, 成为可拓学研究的生力军, 推动可拓学的完善和发展. 我们也希望广大读者对本丛书提出宝贵意见, 为可拓学的建设添砖加瓦.

中国人工智能学会可拓工程专业委员会主任
国家级有突出贡献的专家
新学科可拓学的创立者
蔡　文

2002.6

《可拓学丛书》前言

“可拓学”是以蔡文教授为首的我国学者们创立的新学科，它用形式化的模型，研究事物拓展的可能性和开拓创新的规律与方法，并用于处理矛盾问题.

经过可拓学研究者们多年的艰苦创业、共同奋斗，可拓学已初具规模，包括可拓论、可拓方法、可拓工程等. 在理论和方法研究上取得了创新性、突破性的研究成果，在实际应用中，具有多领域、多类型的成功事例. 可拓学及其应用已引起国内外学术界的广泛关注，具有一定的影响. 其主要成果如下：

★ 可拓论　包括基元理论、可拓集合理论和可拓逻辑.

基元理论提出了描述事、物和关系的基本元 ——“物元”、“事元”和“关系元”，讨论了基元的可拓性和可拓变换规律，研究了定性与定量相结合的可拓模型. 提供了描述事物变化与矛盾转化的形式化语言. 基元理论为知识表示提供了新的形式化工具，可拓模型为人工智能的问题表达提供了定性与定量相结合的模型，对人工智能的发展有重要的意义.

可拓集合论是传统集合论的一种开拓和突破. 它是描述事物“是”与“非”的相互转化及量变与质变过程的定量化工具，可拓集合的质变域和关联函数使可拓集合具有层次性与可变性，从而为研究矛盾问题，发展定量化的数学方法 —— 可拓数学和可拓逻辑奠定基础.

可拓逻辑是研究化矛盾问题为不矛盾问题的变换和推理规律的科学，它是可拓学的逻辑基础.

★ 可拓方法　是可拓论应用于实际的桥梁. 在可拓学研究过程中提出了基于可拓论的多种可拓方法，如发散树、分合链、相关网、蕴含系、共轭对等方法；优度评价、真伪信息判别等评价判别方法；基本变换、复合变换和传导变换等可拓变换方法；菱形思维方法及转换桥方法等综合方法.

★ 可拓工程　将可拓方法应用于工程技术、社会经济、生物医学、交通环保等领域，与各学科、各专业的方法和技术相结合，发展出各领域的

应用技术, 统称为 “可拓工程”. 可拓工程研究的基本思想是用形式化的方法处理各领域中的矛盾问题, 化不相容为相容, 化对应为共存. 近年来, 可拓学在计算机、人工智能、检测、控制、管理和决策等领域进行的应用研究取得了良好的成绩. 实践证明, 可拓学的发展及应用, 具有广阔的前景.

《可拓学丛书》的出版, 总结了多年来可拓学在理论和应用上的研究成果, 这对于可拓学的应用和普及具有重要的意义. 它将推动可拓学研究的深入和发展. 虽然可拓学研究目前已经取得了初步的成绩, 但是还有许多工作要做, 也可能遇到各种各样的困难和挫折. 尽管科学的道路是不平坦的, 但前途是光明的. 特赋诗一首以祝贺《可拓学丛书》的出版:

人工智能天地广,
可拓工程征途长.
中华学者勇创新,
敢教世界看东方.

中国人工智能学会荣誉理事长
《可拓学丛书》编委会主任
涂序彦
2002.6

前　言

数据挖掘研究方兴未艾, 可拓学的应用日渐发展.

数据挖掘研究从数据中挖掘有用的知识; 可拓学研究处理矛盾问题的理论与方法, 而其工具之一是可拓变换, 决策者可以借助可拓变换使矛盾问题转化. 在数据库或数据仓库中, 数据的变化可以反映出可拓变换的作用. 如果能从过去的数据中找到基于可拓变换的知识, 就可以利用它们辅助今天的决策. 相关行业和地区的决策者可以借助这些知识提出解决矛盾问题的策略, 这对社会和经济的发展具有实用价值.

基于可拓变换的知识称为可拓知识, 挖掘各种可拓知识是可拓数据挖掘的重要任务, 这是一片尚待耕耘的领域. 耕耘这片领域的理论依据是以可拓集为基础的可拓论, 使用的是现有数据挖掘与可拓方法结合的可拓数据挖掘方法, 应用的对象是经济、管理、营销、策划、医学、设计等多个领域.

作者从多年的研究实践中体会到这项研究工作的价值, 利用已有的数据挖掘知识和可拓学知识撰写了本专著, 提出了可拓数据挖掘方法. 应该说, 研究成果是初步的. 撰写本书的目的是为有志在这片土地耕耘的学者提供一块敲开可拓数据挖掘大门的敲门砖, 冀求引出有识之士高水平的研究成果.

本书共分 5 章: 第 1 章介绍可拓数据挖掘提出的背景、研究的内容、理论依据、应用方法和发展前景; 第 2 章介绍可拓集与关联函数的基本理论; 第 3 章介绍可拓数据挖掘的基本知识; 第 4 章介绍挖掘基于可拓集的可拓分类知识; 第 5 章介绍挖掘传导知识.

本书是国家自然科学基金项目 “获取变化知识的可拓数据挖掘理论、方法及其实证研究 (70671031)” 和广东省自然科学基金项目 “基于可拓论的信息–知识–策略形式化体系研究 (05001832)” 的研究成果. 作者衷心感谢国家自然科学基金委员会管理科学部多年来对可拓学研究工作的大力支持! 衷心感谢广东省自然科学基金委员会、广东省教育厅和广东工业大学对本研究工作的支持!

由于作者水平有限, 疏漏之处在所难免, 恳请读者不吝批评赐教.

作者谨识

2008.3

目　录

第1章 绪 论

在实际工作中, 要处理各种各样的矛盾问题, 通过可拓变换, 可以使“不是” 变为 “是”, “不行” 变为 “行”[1]. 例如, 通过吃药, 会使病人从有病变为无病, 处方不同, 病人病情的变化会不同, 所属疾病的类型和程度的变化也不同; 在化学实验中, 改变不同的配方, 会得到不同的结果; 在经济活动中, 对银行的利率采用不同的加息措施, 经济从过热转化为不过热的程度和效果会不同; 在市场营销中, 决策者采用不同的措施, 对产品从滞销变为畅销的作用也不同 …… 在这些活动中, 变换的知识起着重要的作用. 由于计算机技术的发展, 在上述活动过程中, 积累了大量的数据. 如何从这些变化的数据中, 挖掘出有用的知识, 进而为解决矛盾问题服务, 这为数据挖掘的研究提出了重要的课题.

1.1 数据挖掘与知识发现

1.1.1 知识发现过程

知识发现 (knowledge discovery in database, KDD) 是从数据中发现有用知识的整个过程 [2]. 数据挖掘是 KDD 过程中的一个特定步骤, 它用专门算法从数据中抽取模式 (pattern)[3].

KDD 是从数据集中识别出有效的、新颖的、潜在有用的以及最终可理解的模式的高级处理过程, 其中, 数据集是事实 F(数据库元组) 的集合; 模式是用语言 L 表示的表达式 E, 它所描述的数据是集合 F 的一个子集 FE, 它比枚举所有 FE 中元素更简单, 称 E 为模式, 发现的模式有一定的可信度, 应该是新的, 将来有实用价值, 能被用户所理解.

根据文献 [3], KDD 的过程如图 1.1 所示.

KDD 过程可以概括为三部分: 数据准备 (data preparation)、数据挖掘 (data mining) 及结果的解释和评估 (interpretation & evaluation).

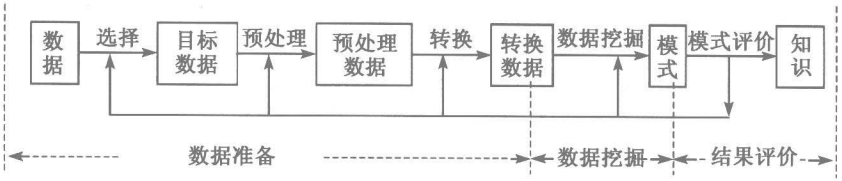

图 1.1　KDD 过程图

1) 数据准备

数据准备可分为三个子步骤：数据选择 (data selection)、数据预处理 (data preprocessing) 和数据转换 (data transformation).

数据选择的目的是确定发现任务的操作对象, 即目标数据 (target data), 它根据用户的需要, 从原始数据库中选取一组数据. 数据预处理一般包括消除噪声、推导计算缺值数据、消除重复记录等. 数据转换的主要目的是完成数据类型转换 (如把连续值数据转换为离散型数据, 以便于符号归纳; 或是把离散型数据转换为连续值型数据, 以便于神经网络计算), 进行属性约简, 即从初始属性中找出真正有用的属性, 删除无用属性, 以减少数据挖掘时要考虑的属性个数.

2) 数据挖掘

数据挖掘阶段首先要确定挖掘的任务或目的, 如数据分类、聚类、关联规则发现或序列模式发现等. 确定了挖掘任务后, 就要决定使用什么样的挖掘算法. 选择实现算法需要考虑两个因素：一是不同的数据有不同的特点, 因此需要用与之相关的算法来挖掘; 二是用户或实际运行系统的要求, 有的用户可能希望获取描述型的 (descriptive)、容易理解的知识 (采用规则表示的挖掘方法显然要好于神经网络之类的方法), 而有的用户只是希望获取预测准确度尽可能高的预测型 (predictive) 知识. 选择了挖掘算法后, 就可以实施数据挖掘操作, 获取有用的模式.

3) 结果的解释和评估

数据挖掘阶段发现出来的模式, 经过评估, 可能存在冗余或无关的模式, 这时需要将其剔除; 也有可能模式不满足用户要求, 这时则需要回退到发现过程的前面阶段, 如重新选取数据、采用新的数据变换方法、设定新的参数值, 甚至换一种挖掘算法等. 另外, KDD 由于最终是面向人类用

户的, 因此可能要对发现的模式进行可视化, 或者把结果转换为用户易懂的另一种表示, 如把分类决策树转换为 “if···then···” 规则.

数据挖掘仅仅是整个过程中的一个步骤, 数据挖掘质量的好坏有两个影响要素: 一是所采用的数据挖掘技术的有效性, 二是用于挖掘的数据的质量和数量 (数据量的大小). 如果选择了错误的数据或不适当的属性, 或对数据进行了不适当的转换, 则难以挖掘到好的结果.

整个挖掘过程是一个不断反馈的过程. 例如, 如果用户在挖掘过程中发现选择的数据不太好, 或使用的挖掘技术产生不了期望的结果, 这时用户就需要重复前面的过程, 甚至从头开始.

可视化技术在数据挖掘的各个阶段都发挥着重要的作用, 特别是在数据准备阶段, 用户可能要使用散点图、直方图等统计图形的可视化技术来显示有关数据, 以期对数据有一个初步的了解, 从而为更好地选取数据打下基础. 在挖掘阶段, 用户则要使用与该领域问题有关的可视化工具. 在表示结果阶段, 则可能要用到可视化技术以使得发现的知识更易于理解.

1.1.2　数据挖掘的对象

数据挖掘的对象主要是关系数据库和数据仓库, 这是典型的结构化数据. 随着技术的发展, 数据挖掘对象逐步扩大到半结构化或非结构化数据, 如文本数据、图像和视频数据以及 Web 数据等.

1. *关系数据库*

目前建立的数据库都是关系数据库, 数据仓库的数据存储仍然是关系数据库. 数据挖掘方法也主要是研究数据库中属性之间的关系, 挖掘出多个属性取值之间的规则. 数据库的特点有

1) 数据的动态性

数据的动态变化是数据库的一个主要特点. 由于数据的存取和修改, 使数据的内容经常发生变化, 这就要求数据挖掘方法能适应这种变化. 渐增式数据挖掘方法就是针对数据变化后挖掘的规则知识能满足变化后的数据库内容的方法.

2) 数据的不完全性

这主要反映在数据库中记录的域值丢失或不存在 (空值), 这种不完

全数据给数据挖掘带来了困难. 为此, 必须对数据进行预处理, 填补该数据域的可能值.

3) 数据噪声

由于数据录入等原因造成错误的数据, 即数据噪声. 挖掘含噪声的数据会影响获取模式的准确性, 并增加数据挖掘的困难度. 因此, 在数据挖掘中要考虑噪声的影响, 利用概率方法排除这些噪声. 目前有些专家专门研究含噪声数据的挖掘方法.

4) 数据冗余性

数据冗余性表现为同一信息在多处重复出现. 函数依赖是一个通常的冗余形式. 冗余信息可能造成错误的数据挖掘, 至少有些挖掘的知识是用户不感兴趣的. 为避免这种情况的发生, 数据挖掘时, 需要知道数据库中有哪些固有的依赖关系.

5) 数据稀疏性

数据稀疏性表现为多维数据空间中存在大量稀疏数据, 稀疏数据会使数据挖掘丢失有用的模式.

6) 海量数据

数据仓库中数据在不断增长, 已出现很多海量数据仓库. 数据挖掘方法需要逐步适应这种海量数据和迅速增长的数据挖掘, 如建立有效的索引机制和快速查询方法、采用分布式计算技术等.

2. 文本

文本是以文字串形式表示的数据文件. 文本分析包括: 关键词或特征提取、相似检索、文本聚类和文本分类等.

1) 关键词或特征提取

一篇文本中, 标题是该文本的高度概括, 标题中的关键词是标题的核心内容, 关键词的提取对于掌握该文本的内容至关重要.

文本中的特征, 如人名、地名、组织名等是某些文本中的主体信息, 特征提取对掌握该文本的内容很重要.

2) 相似检索

文本中的关键词的相似检索是了解文本内容的一种重要方法, 如 “专家系统” 与 “人工智能” 两个关键词是有一定联系的, 研究专家系统的文

本一定属于人工智能的研究领域.

3) 文本聚类

对于文本标题中关键词 (主题字) 的相似匹配是对文本聚类的一种简单方法. 定义关键词的相似度, 将便于文本的简单聚类, 使类中的文本均满足关键词的相似度, 使类间的文本的关键词一定超过相似度.

4) 文本分类

将文本分类到各文本类中, 一般需要采用一个算法. 这些算法包括分类器算法、近邻算法等. 这需要按文本中的关键字或特征的相似度来区分.

3. 图像与视频数据

图像和视频数据是典型多媒体数据. 数据以点阵信息及帧形式存储, 数据量很大. 图像与视频的数据挖掘包括: 图像与视频特征提取, 基于内容的相似检索, 视频镜头的编辑与组织等.

1) 基于内容的相似检索

根据图像、视频特征的分布、比例等进行基于内容的相似检索, 可以将图像和视频数据进行聚类以及分类, 也能完成对新图像或视频的识别, 如对遥感图像或视频的识别. 这种应用非常广泛, 如森林火灾的发现与报警, 河流水灾的预报等.

2) 图像与视频特征提取

图像与视频数据特征有颜色、纹理和形状等. 这些特征提取在基于内容的相似检索中有较好的应用. 房屋的形状及颜色等, 需要从大量图像和视频数据中提取, 如 (海水, 颜色, 蓝色)、(海滩, 颜色, 黄色) 等.

3) 视频镜头的编辑与组织

镜头代表一段连续动作 (视频数据流). 典型的镜头编辑如足球赛的射门、某段新闻节目等, 需要在冗长的视频数据流中进行自动裁取.

经过编辑的镜头, 按某种需要重新组织, 将形成特定需求的新视频节目, 如足球射门集锦, 某个新闻事件的连续报道等.

4. Web 数据

随着 Internet 的发展和普及、网站数目的迅速增长以及入网人员的迅速增加, 网络数据量呈指数增长. Web 数据挖掘已成为新课题. Web 数

据挖掘的特点有

1) 异构数据集成和挖掘

Web 上每一个站点是一个数据源, 各数据源都是异构的, 形成了一个巨大的异构数据库环境. 将这些站点的异构数据进行集成, 给用户提供一个统一的视图, 才能在 Web 上进行数据挖掘.

2) 半结构化数据模型抽取

Web 上的数据非常复杂, 没有特定的模型描述. 虽然每个站点上的数据是结构化的, 但各自的设计对整个网络是一个非完全结构化的数据, 称为半结构化数据.

对半结构化数据模型的查询和集成, 需要寻找一种半结构化模型抽取技术来自动抽取各站点的数据.

XML 是一种半结构化的数据模型, 容易实现 Web 中的信息共享与交换.

采用 "实时建议" 技术, 能够根据用户以往的浏览行为来预测该用户以后的浏览行为, 从而为用户提供个性化的浏览建议.

总之, Web 数据挖掘正在逐步形成热点.

1.1.3　数据挖掘的任务

现有数据挖掘的任务有六项：关联分析、时序模式、聚类、分类、偏差检测、预测.

1. 关联分析

关联分析是从数据库中发现知识的一类重要方法. 若两个或多个数据项的取值之间重复出现且概率很高时, 它就存在某种关联, 可以建立起这些数据项的关联规则.

例如, 买面包的顾客有 90%的人还买牛奶, 这是一条关联规则. 若商店中将面包和牛奶放在一起销售, 将会提高他们的销量.

在大型数据库中, 这种关联规则是很多的, 需要进行筛选, 一般用 "支持度" 和 "可信度" 两个阈值来淘汰那些无用的关联规则.

2. 时序模式

通过时间序列搜索出重复发生概率较高的模式. 这里强调时间序列

的影响. 例如, 在所有购买了激光打印机的人中, 半年后 80%的人再购买新硒鼓, 20%的人用旧硒鼓装碳粉; 在所有购买了彩色电视机的人中, 有 60%的人再购买 VCD 产品.

在时序模式中, 需要找出在某个最小时间内出现比率一直高于某一最小百分比 (阈值) 的规则. 这些规则会随着形式的变化作适当的调整.

时序模式中, 一个有重要影响的方法是 "相似时序". 用 "相似时序" 的方法, 要按时间顺序查看时间事件数据库, 从中找出另一个或多个相似的时序事件. 例如, 在零售市场上找到另一个有相似销售的部门, 在股市中找到有相似波动的股票.

3. 聚类

数据库中的数据可以划分为一系列有意义的子集, 即类. 简单地说, 在没有类的数据中, 按 "距离" 概念聚集成若干类. 在同一类别中, 个体之间的距离较小, 而不同类别上的个体之间的距离偏大. 聚类增强了人们对客观现实的认识, 即通过聚类建立宏观概念, 如将鸡、鸭、鹅等都聚类为家禽.

聚类方法包括统计分析方法、机器学习方法、神经网络方法等.

在统计分析方法中, 聚类分析是基于距离的聚类, 如欧氏距离、汉明距离等. 这种聚类分析方法是一种基于全局比较的聚类, 它需要考察所有的个体才能决定类的划分.

在机器学习方法中, 聚类是无导师的学习. 在这里距离是根据概念的描述来确定的, 故聚类也称概念聚类, 当聚类对象动态增加时, 概念聚类则称为概念形成.

在神经网络中, 自组织神经网络方法用于聚类, 如 ART 模型、Kohonen 模型等, 这是一种无监督学习方法. 当给定距离阈值后, 各样本按阈值进行聚类.

4. 分类

分类是数据挖掘中应用最多的任务. 分类是在聚类的基础上, 对已确定的类找出该类别的概念描述, 它代表了这类数据的整体信息, 即该类的内涵描述. 一般用规则或决策树模式表示. 该模式能把数据库中的元组影