可拓学丛书

可拓数据挖掘方法及其计算机实现

杨春燕 李小妹
陈文伟 蔡 文 著

广东高等教育出版社
Guangdong Higher Education Press

可拓学丛书

可拓数据挖掘方法及其计算机实现

杨春燕　李小妹
陈文伟　蔡　文　著

国家自然科学基金资助项目
广东省普通高校人文社会科学研究重点项目

广东高等教育出版社
·广州·

内容简介

可拓数据挖掘方法是以可拓集为集合论基础，从数据库或知识库中获取基于变换的知识（简称可拓知识）的有效方法. 本书依据可拓数据挖掘的基本原理和方法，结合最新的人工智能理论与工具，研究了可拓数据挖掘方法的可操作性，包括基于数据库的可拓分类知识挖掘、基于数据库的传导知识挖掘、基于数据库的可拓聚类知识挖掘、基于知识库的可拓知识挖掘等内容，并利用案例介绍了各种方法的计算机实现.

本书将可拓数据挖掘方法与应用相结合，分析透彻，可操作性强，适合高等院校智能科学、计算机科学、管理科学与工程等领域的师生，企事业的工程技术人员和管理决策人员阅读.

图书在版编目（CIP）数据

可拓数据挖掘方法及其计算机实现/杨春燕等著. —广州：高等教育出版社，2010.8

（可拓学丛书）

ISBN 978-7-5361-3973-2

Ⅰ. ①可… Ⅱ. ①杨… Ⅲ. ①数据采集-研究 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2010）第 174789 号

广东高等教育出版社 出版

地址：广州市天河区林和西横路

邮编：510500 电话：87557232

http：//www.gdgjs.com.cn

佛山市浩文彩色印刷有限公司印刷

*

2010 年 8 月第 1 版 开本：32 开（890 mm×1 240 mm）

2010 年 8 月第 1 次印刷 印张：8.25

印数：1—2 000 字数：280 千字

定价：32.00 元

（如有印装质量问题，我社负责调换）

《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

可拓数据挖掘是可拓学和数据挖掘结合的产物，它研究用可拓学的理论和方法，应用数据挖掘技术去挖掘数据库中与解决矛盾问题的变换有关的知识.

研究显示，把可拓学与数据挖掘相结合，将发展现有的数据挖掘理论和技术，产生新的可拓数据挖掘理论与技术. 将该技术应用于市场营销、客户关系管理、金融证券、电信、医疗等领域数据的挖掘研究，可为解决其中的矛盾问题提供有效的决策支持.

可拓数据挖掘的重点是挖掘基于变换的知识（可拓知识），它是在数据挖掘的基础上，利用可拓变换，获取可拓知识，包括可拓分类知识、传导知识、可拓聚类知识，以及其他关于变换的知识，为解决矛盾问题提供依据，这是利用数据挖掘技术解决矛盾问题的新思路.

本书第 1 章是绪论，简要介绍可拓数据挖掘方法及其计算机实现研究的意义、价值、现状、主要内容及发展前景；第 2 章介绍与本书相关的知识，包括可拓变换、可拓信息与可拓知识等；第 3 章介绍基于数据库的可拓分类知识挖掘；第 4 章介绍基于数据库的传导知识挖掘；第 5 章介绍基于数据库的可拓聚类知识挖掘；第 6 章介绍基于知识库的可拓知识挖掘. 各部分内容都以若干案例来帮助读者理解. 我们期望高等院校和科研单位的相关专业的教学科研人员，能将这些方法与自己的研究领域相结合，在书中介绍的实验软件的基础上，研制出可供相关专业应用的应用软件.

本书是作者承接的国家自然科学基金项目（70671031）和广东省普通高校人文社会科学研究重点项目（06ZD63008）的研究成果，作者冀求以此拙作作为引玉之砖，以使更多相关领域的学者利用可拓数据挖掘方法去获取所在领域的可拓知识. 同时，我们也希望这本书能成为可拓学与计算机、人工智能结合的桥梁.

作者感谢国家自然科学基金委员会和广东省教育厅对我们的研究工作给予的大力支持！感谢广东工业大学为我们提供的良好科研环境！感谢李承晓、邵松华、方卓君、廖美东、何小龙等同学为本书的写作收集和整理资料、参加软件研制工作；感谢本研究团队的李卫华教授对本书写作的支持！感谢广东高等教育出版社对本书的出版所付出的辛勤工作！

由于作者才疏学浅，疏漏乃至错误之处在所难免，恳请读者批评指正！

作者谨识

2010.4.2

目　　录

第1章 绪 论

可拓学是以形式化的模型，探讨事物拓展的可能性以及开拓创新的规律与方法，并用于解决矛盾问题的学科[1-6]. 所谓矛盾问题，就是指在现有条件下无法实现人们要达到的目标的问题. 可拓学研究如何通过可拓变换，使矛盾问题变为不矛盾问题. 可拓学的基本理论与方法和各领域的知识相结合，拓展了该领域的理论，也产生了处理该领域矛盾问题的可拓工程方法. 目前，这些工作已经取得了很多成果[7].

可拓数据挖掘[7-9]是可拓学和数据挖掘结合的产物，它研究用可拓学的理论和方法，去挖掘数据库或知识库中与解决矛盾问题的变换有关的知识. 本书讨论利用可拓学的基本理论和方法，去挖掘各个领域的数据库或知识库中所积累的基于变换的知识.

1. 研究意义与价值

一项重大政策的实施，会影响社会经济的发展，从而使统计数据产生变化；反之，从社会经济的统计数据库中，可以挖掘政策的作用规律. 在很多行业中，需要人们考虑研究对象的各特征之间，以及各研究对象之间存在什么样的相关关系，采用什么变换去处理矛盾问题，了解实施某个变换以后产生怎样的效应，等等. 例如，当经济过热或经济衰退时，银行采取加息或减息去处理问题. 那么，能否知道这些措施的作用效果，以便对变化的数据、范围和影响速度有一定的认识？相反，能否从数据库积累的大量数据中找到与这些措施的实施有关的知识，来帮助今后制定相应的决策？例如何时加息为宜，幅度多大为好等等. 这类问题比比皆是.

另一方面，要使计算机具有较高的智能水平，必须研究解决矛盾问题的策略生成理论与方法，研究矛盾问题智能化处理的规律和技术. 解决矛盾问题的关键是变换，必须研究如何寻找变换，分析变换的作用，从数据库中获取变换对数据变化的作用有关的知识，才能为生成处理矛

盾问题的策略提供依据.

基于上述两个原因，人们必须研究适合于变换下数据变化规律的新的数据挖掘理论和方法.

随着计算机软硬件技术的迅速发展，各行各业都逐步建立了自己的数据库，所收集的数据量以每年翻一番的速度增长. 大量积累的数据之中隐藏着丰富的知识，运用这些知识可以对未来的工作有所指导和帮助. 然而，如此大的数据量和增长速度，使用人工分析去获得潜在知识是无能为力的. 数据挖掘作为信息技术发展的一项关键技术，在市场营销方面已经产生了巨大的价值.

但是，现有的数据挖掘理论与方法是挖掘在不变化条件下的知识，而对于如何从浩瀚的数据中寻找变换导致数据变化的规律，尚未见到有关的研究.

可拓学研究通过变换处理矛盾问题的规律和方法，其集合论基础是可拓集论. 从可拓集的观点看，数据库中的一个事项就是一个 n 维基元，一个数据表就是一个基元域. 利用关联函数可以计算数据和事项的关联度，去表示事物符合某一要求的程度，通过可拓规则研究变换对数据作用的规律，利用可拓集合和关联函数的性质及运算，通过可拓推理确定可拓变换的变源和内涵.

可拓数据挖掘从理论上把可拓集理论和可拓逻辑应用于数据挖掘中，把数据库和数据仓库与可拓集的论域对应起来，把数据与基元对应起来，从而可以把可拓集理论与数据挖掘这一领域相结合，形成挖掘“可拓知识”的基本理论. 在方法上，把以基元为逻辑细胞的形式化体系与数据库和数据仓库结合起来，形成适合于数据“变换”的知识表示方法，利用可拓推理和关联函数为工具，建立一套适合于挖掘可拓知识的可拓数据挖掘方法，进而开发获取“可拓知识”的实验软件.

本书探讨把用形式化模型处理矛盾问题的可拓学理论与方法和数据挖掘技术相结合，利用可拓方法挖掘数据库或知识库中的“可拓知识”和“变换的作用”，进而挖掘数据库中由于政策的变化（用可拓变换形式化描述）导致数据变化的规律，以及知识库中由于变换的作用导致知识的变化的规律，从而为决策者提供决策参考依据，是一项很有价值的开创性工作. 这对于发展数据挖掘的理论和方法具有一定的科学意义，

对于社会经济各行业（如金融、税务、房地产等）寻找处理问题的策略和分析变换的作用有较大的实用价值.

2. 国内外研究现状

随着信息技术的发展，信息处理模式已由事务处理、分析处理发展为知识发现（KDD）. 知识发现是一个多步骤的对大量数据进行分析的过程，包括数据预处理、模式提取、知识评估及过程优化. 数据挖掘（DM，也称数据开采）是大型数据库中知识发现的一个关键步骤，SAS软件研究所对它下的定义是："按照既定的业务目标，对大量的企业数据进行探索，提出隐藏于其中的规律，并进一步模型化的先进、有效的方法." 但是，"有关数据挖掘的理论基础研究还没有成熟. 坚实的和系统的理论基础对于数据挖掘非常重要，因为它给数据挖掘技术的开发、评价和实践提供了一致的框架"[10].

数据挖掘[11][12]自20世纪80年代中期提出以来，得到了迅速的发展. 从数据挖掘的任务考虑，主要从事关联规则、时序模式、聚类数据、分类数据、偏差分析和预测数据等方面有用的知识挖掘. 从挖掘方法和技术考虑，目前使用较多的有归纳学习方法（如信息论方法和集合论方法）、仿生物技术（如神经网络方法和遗传算法）、公式发现、统计分析方法、模糊数学方法、可视化技术等.

数据挖掘是建立在数据库基础之上的自动数据分析技术，它能够在海量的数据中快速地寻找到一些十分有价值、有意义的数据间的特定关系并产生新的知识，这些知识可以为管理者提供有力的决策支持. 但存在如下问题：

（1）现有的数据挖掘技术基本上是在静态数据中获取知识，由于数据库的静态性，挖掘出的知识也属于静态知识，未能反映知识的动态特性，没有考虑变换及变换的作用、可拓知识、变换蕴含式等的挖掘，而这些是解决矛盾问题所需要的.

（2）对关联规则技术的研究主要集中在先从数据集中找出出现最为频繁的数据项（组合），然后寻找出频繁集中某些数据项对其他数据项的影响作用. 然而，只用出现的频率来衡量数据项的重要性，忽略了出现频率的变化、数据项的数量、价值、利润等更有实际意义的衡量指标，

而由于“关联规则”类问题固有的计算复杂度，现有的高速算法不能支持较为复杂的衡量指标. 此外，现有的算法所能挖掘出的数据项间相互影响的规则只停留在已发生的数据项集、寻找已出现的联系，而不能提供更深层次的、变化的各数据项间关系的刻画，因此对于决策支持的力度有限.

（3）分类技术是指预测出一个新的数据记录的类别，其中预测的依据是由训练集数据中分析得来的数据各属性与已知类别间的函数关系. 然而，现有算法的可适性较差，即添加任何属性或扩大某个属性的取值域都要重新训练，找到新的函数关系，费时费力，不适应递增的渐进学习，对决策的支持有局限性.

20 世纪 60 年代提出的模糊集[13]研究了量值的模糊化问题，80 年代提出的粗糙集[14]研究了数据中属性约简的方法. 但无论是经典集、模糊集或者粗糙集，它们处理的都是静态的知识，不是变换下的知识，而可拓集[15]则是研究变换下事物性质的变化规律，从定性和定量的角度描述量变和质变.

可拓数据挖掘重点挖掘基于变换的知识（可拓知识），它是在数据挖掘的基础上，增加可拓变换，获取可拓知识，为解决矛盾问题提供依据，这是利用数据挖掘技术解决矛盾问题的新思路.

研究显示，把可拓学与数据挖掘相结合，将发展现有的数据挖掘理论和技术，产生新的可拓数据挖掘理论与技术. 将该技术应用于市场营销、客户关系管理、金融证券、电信、医疗等领域数据的挖掘研究，可为解决其中的矛盾问题提供有效的决策支持.

3. 本书主要内容

本书的框架结构如图 1.1.

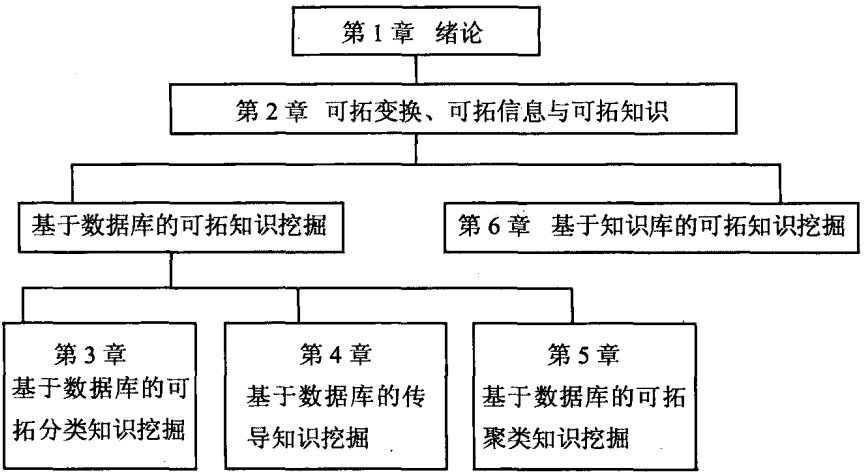

图 1.1 本书的框架结构图

第2章是全书的基础，重点介绍在可拓数据挖掘中涉及的一些有关可拓学的基本知识，包括：

（1）基元与复合元的概念和结构形式；

（2）可拓变换的一般定义、类型、性质以及基本运算；

（3）信息的基元表示方法、拓展信息的定义及其表示方法、可拓信息的定义及其表示方法；

（4）传导变换的一般概念、传导效应的定义及计算方法、共轭部的变换和共轭变换；

（5）可拓知识的定义及其类型.

第3章详细介绍了基于数据库的可拓分类知识挖掘方法及其计算机实现，包括：

（1）分类方法综述；

（2）可拓分类的集合论基础——可拓集简介、关联函数的选取与综合关联函数的构造方法、可拓分类方法的一般步骤；

（3）可拓分类知识的含义及表示方法、支持度和可信度的计算方法、可拓分类知识的挖掘方法；

（4）以某企业的产品销售问题为例，介绍可拓分类知识挖掘的计算机实现.

第4章详细介绍了基于数据库的传导知识挖掘方法及其计算机实现，包括：

（1）传导特征和传导对象的概念、传导度的一般定义及各种情形下的计算方法；

（2）传导知识的类型；

（3）传导知识的挖掘方法，包括：从某一对象的多个特征量值的数据库中挖掘传导知识的方法，从多个对象的统一特征量值的数据库中挖掘传导知识的方法，以及从多对象多特征量值的数据库中挖掘传导知识的方法；

（4）以成品油价税费改革对股票市场的影响为例，介绍传导知识挖掘的计算机实现.

第5章详细介绍了基于数据库的可拓聚类知识挖掘方法，包括：

（1）聚类方法综述；

（2）单评价特征和多评价特征情况下关联度的计算方法、可拓聚类方法的一般步骤；

（3）可拓聚类知识的表示方法、可拓聚类知识挖掘的一般步骤；

（4）以服用某种保健品对女性血压的影响为例，介绍了可拓聚类知识挖掘过程.

第6章详细介绍了基于知识库的可拓知识挖掘方法及其计算机实现，包括：

（1）基于拓展规则的知识——拓展型知识的表示方法、基于拓展型知识的可拓知识及其获取方法；

（2）决策树知识简介、决策树的构成方法、基于决策树知识的可拓知识及其获取方法，并以大学英语四级通过情况为例介绍这种可拓知识挖掘的计算机实现；

（3）基于本体的可拓知识链的定义、获取方法及实例，解决矛盾问题的可拓知识，基于属性约简变换和数据挖掘变换的可拓知识.

4. 可拓数据挖掘的发展前景

随着信息技术的迅速发展，管理信息系统、互联网、数据挖掘、知识管理等正在不断积累越来越多的数据、信息和知识. 在这种情况下，

企业更需要深层的、实用性强的知识辅助决策. 例如，在市场竞争日趋激烈的今天，客户成为重要资源，变换的知识有助于将初次注册用户和即将流失的客户转化为忠诚客户，降低客户保持和新客户开发的成本；在信用风险分析中，既要识别高风险客户，也要采取措施促使有欺诈动机的客户停止行动，可拓分类方法及其相关的知识就大有用武之地；在新产品研发中，蕴含性分析的知识有助于及早发现产品的趋势，识别客户的潜在需求；在业务流程优化中，可拓数据挖掘有助于找出影响效率的瓶颈环节并采取转化措施；在医疗行业，变换的知识可帮助医生及早发现症状的根本变化，识别效用最好的方案以提高治疗效果；在市场营销中，变换的知识对开发市场具有指导意义. 总之，在事物类别转化、发现问题的根本原因、识别潜在的变换知识等方面，可拓数据挖掘都可以发挥作用. 因此，可拓数据挖掘具有广阔的应用前景.

可拓论和可拓方法与信息、管理等领域交叉融合产生的可拓工程已经取得了初步的成果. 目前，在可拓数据挖掘领域，一些专家已经利用可拓学原理和方法，从可拓数据挖掘的理论、方法、算法、应用方案、技术改进等方面对上述数据挖掘中存在的部分问题进行了研究，并取得了一定的成果. 随着研究的深入和研究力度加强，有望产生有更大应用价值的研究成果.

随着经济全球化的推进，知识经济的步伐也大大加快，环境的多变促使了信息和知识的更新周期缩短，创新和解决矛盾越来越成为各行各业的重要工作. 因此，如何挖掘变换的知识就成为数据挖掘研究的重要任务.

目前对可拓数据挖掘的研究重点限于对关系数据库或规则知识库. 实际上，在研究文本数据、图象与视频数据、Web 数据等时，也都应该考虑变换对数据的影响，也是可拓数据挖掘应该涉足的领域. 对数据仓库中可拓知识的挖掘也刚刚涉足，还有待于今后进一步深入研究.

第2章　可拓变换、可拓信息与可拓知识

可拓数据挖掘主要的研究内容是研究如何获取基于可拓变换的知识，简称可拓知识. 而知识是由信息构成的，基于可拓变换的信息，称为可拓信息. 可拓变换的对象可以是基元或复合元，也可以是关联准则或论域，为此，本章首先介绍基元、复合元和可拓变换的基本知识，进而介绍可拓信息与可拓知识的概念.

2.1　基元与复合元

2.1.1　基元

大千世界由万物构成，物与物的相互作用就是事，物与物、物与事、事与事之间存在各种关系，物、事和关系千变万化，形成了五彩缤纷的天地人世.

为了让计算机能描述各种现象及其变化，必须研究简练、方便的形式化模型. 经过多年的探索，我们建立了可拓模型，它以基元为逻辑细胞，包括物元、事元和关系元. 描述物的是物元，描述事的是事元，描述关系的是关系元. 对于复杂的现象，可以用它们复合而成的复合元或基元的运算式来表示.

1. **物元**（matter-element）

定义 2.1　以物 O_m 为对象，c_m 为特征，O_m 关于 c_m 的量值 v_m 构成的有序三元组

$$M=(O_m, c_m, v_m)$$

作为描述物的基本元，称为一维物元，O_m，c_m，v_m 三者称为物元 M 的三要素，其中 c_m 和 v_m 构成的二元组（c_m，v_m）称为物 O_m 的特征元.

为方便起见，把物元的全体记为$\pounds(M)$，物的全体记为$\pounds(O_m)$，特