规则挖掘技术

张德干 王晓晔 ◎ 著

科学出版社
www.sciencep.com

规则挖掘技术

张德干　王晓晔　著

科学出版社

北京

内 容 简 介

规则挖掘技术是指从数据库中抽取隐含的、潜在的、先前未知的、有用的知识和规则的一门交叉学科技术。它受多个学科的影响，同时它又对多个学科的发展、应用产生积极而深远的影响，具有十分重要的促进作用。本书涉及的内容有规则挖掘技术概论、具有冗余约简能力的规则挖掘机制、分明关系约束的格上规则挖掘方法、基于包含度的决策树中规则挖掘方法、基于时间序列的规则挖掘方法、规则挖掘过程中的分类技术、应用案例等内容。

本书介绍的规则挖掘技术新颖、涵盖面广、信息量大、实用性强。本书图文并茂，十分方便本科生、研究生、教师学习和参考，也非常方便从事数据挖掘以及相关领域的科研和工程开发技术人员阅读、参考。

图书在版编目(CIP)数据

规则挖掘技术/张德干，王晓晔著．—北京：科学出版社，2008
ISBN 978-7-03-023092-8

Ⅰ．规… Ⅱ．①张… ②王… Ⅲ．数据库系统 Ⅳ．TP311.13

中国版本图书馆 CIP 数据核字（2008）第 157041 号

责任编辑：余 江 于宏丽 / 责任校对：陈玉凤
责任印制：张克忠 / 封面设计：耕者设计工作室

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

骏走印刷厂印刷

科学出版社发行 各地新华书店经销

*

2008 年 12 月第 一 版 开本：B5(720×1000)
2008 年 12 月第一次印刷 印张：13 1/4
印数：1—3 000 字数：247 000

定价：32.00 元

（如有印装质量问题，我社负责调换〈环伟〉）

前　言

规则挖掘(rule mining)是许多传统学科和新兴工程领域相结合而产生的一个年轻而又活跃的前沿技术领域，是多种智能控制系统的重要组成部分。它是指从数据库中抽取隐含的、潜在的、先前未知的、有用的知识或规则的一门交叉学科技术。由于其潜在的理论意义和巨大的应用价值，世界各国都投入了大量的人力、物力和财力，进行广泛深入的研究。就我们所知，到目前为止，规则挖掘技术还有诸多值得研究的方面，而研究的很多成果也还远没有大规模地呈现在如电子商务、机器人、交通管制、军事应用等应用领域。

本书针对规则挖掘技术中的如下几个关键问题进行了研究：规则挖掘机制、规则挖掘方法、规则挖掘过程中的分类、规则挖掘技术的应用案例等。

全书共分9章。其中，第一章综述了规则挖掘技术研究的背景，第二章阐述了规则挖掘的相关技术，第三章研究了一种具有冗余约简能力的规则挖掘机制，第四章研究了分明关系约束的格上规则挖掘方法，第五章研究了基于包含度的决策树中规则挖掘方法，第六章对前两章的两种方法进行了理论分析与比较，第七章研究了基于时间序列的规则挖掘方法，第八章研究了规则挖掘过程中的分类技术，第九章是本书所研究规则挖掘技术的应用案例。

本书除第七章、第八章由王晓晔撰写外，其余各章均由张德干撰写，并由张德干统稿。本书得到国家863计划项目(No. 2007AA01Z188)、国家自然科学基金项目(No. 60773073，No. 60604010)、教育部重点项目(No. 208010)和浙江大学工业控制技术国家重点实验室项目(No. 0708007)的资助。

本书由张桦教授和郑刚教授审阅。

本书在撰写过程中，多位教授和专家学者提出了建设性意见，同时，得到了韩静等同事和张小丽、李林青、凌辰、李淼、胡素蕊等研究生的支持和帮助，在此一并表示衷心的感谢。本书属研究型专著，可供高校研究生、科研人员和工程技术人员参考。书中不当之处，真诚欢迎各位读者批评指正。

作　者

2008年8月

目　录

此为试读，需要完整PDF请访问：www.ertongbook.com

第一章　绪　　论

规则挖掘是许多传统学科和新兴工程领域相结合而产生的一个新的前沿技术领域，是多种智能控制系统的重要组成部分。无论在军事上，还是在民用上，它已发展成为一个十分活跃的热门研究领域，是多学科、多部门、多领域所共同关心的高层次共性关键技术，包括中国在内的众多国家都相继把它列为未来重点发展的对象。作为一种自动化智能信息综合处理技术，它充分利用多源异类信息的互补性和计算机的高速处理与智能判定来提高结果信息的质量。这一技术首先广泛用于军事，并很快推广到自动控制、航空交通管制、遥感测量以及医疗诊断等众多领域。因其潜在的巨大应用价值，世界各国都投入了大量的人力、物力和财力，进行广泛深入的研究。就我们所知，到目前为止，规则挖掘作为一门学科还未形成一套系统而完备的理论，并有诸多值得研究的方面，而研究的很多成果还远没有大规模地呈现在应用领域。

1.1　数据挖掘技术概论

数据挖掘是20世纪90年代兴起的一项新技术，它是知识发现的关键步骤，国内外学术界和企业界都非常重视对数据挖掘技术和软件工具的研究和开发。数据挖掘是多门学科和多门技术相结合的产物，也是一个非常年轻而又活跃的研究领域。在促进数据挖掘诞生、发展、应用的众多原因中，主要有4种，即超大规模数据库的出现、先进的计算机技术、经营管理的实际需要和对这些数据的精深计算能力。从经营管理角度出发，进入21世纪以后，全球经济一体化的进程日益加快，企业所面临的市场竞争压力日趋严重，企业经营管理者特别是决策者希望能够从企业积累的大量历史数据中找到应对日趋严重的竞争压力的良方，希望能够从这些数据中找到经营管理中问题的根本原因，能够快速从大量数据中挖掘出对经营管理有用的信息，以应对瞬息万变的市场压力。因此可以说数据挖掘技术是一个对管理决策者提供决策支持的有力工具。

面对信息社会中数据和数据库的爆炸式增长，人类分析数据和从中提取有用信息的能力远远不能满足实际需要。所以迫切需要一种能够智能地、自动地把数据转换成有用信息和知识的技术和工具。数据库管理系统和人工智能中的机器学习两种技术的发展和结合促成了知识发现(knowledge discovery in database，KDD)这一新技术的产生。1989年8月在美国底特律召开的第11界国际人工智

能联合会议的专题讨论会上首次提出了 KDD。它是一门交叉性学科，内涵极为广泛，理论和技术难度很大，所以针对大型数据库的 KDD 技术一时还难于满足应用的需要。于是 1995 年，在美国计算机年会(ACM)上提出了数据挖掘(data mining)概念。也有一些文献把数据挖掘技术称为知识抽取(knowledge extraction)、数据考古学(data archaeology)、数据捕捞(data dredging)等。多数人认为数据挖掘是 KDD 过程的关键技术(图 1.1)，从而不加区分地使用知识发现和数据挖掘两个术语。

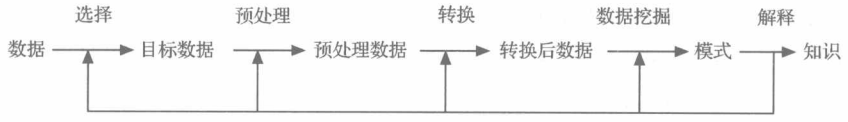

图 1.1 KDD 过程

从技术角度看，数据挖掘是指从数据库中抽取隐含的、潜在的、先前未知的、有用的信息(如知识、规则、约束和规律等)的一个非平凡过程。从广义上理解，数据、信息也是知识的表达形式，但人们更将概念、规则、模式、规律和约束等看成知识。

人们将数据看做形成知识的源泉，原始数据可以是结构化的，如关系数据库中的数据；也可以是半结构化的，如文本、图像数据；甚至是分布在网络上的异构数据。发现知识的方法可以是数学的、非数学的、演绎的和归纳的。发现的知识可以用于信息管理、查询优化、决策支持和过程控制等。它把人们对数据的应用从低层次的简单查询提升到从数据库中挖掘知识，提供决策支持。

1.2 规则挖掘技术的研究意义

分析大量的最新研究文献可知，用于决策的信任度高的规则挖掘问题是多种智能控制领域目前亟待解决的重要课题之一。说它重要，是因为：

(1) 当前众多动态应用场景中，如工控领域中的设备运行状态判断、军事领域中(如美国制定的 DARPA 计划等)的敌我态势评估等，都迫切需要根据当前的多个状态估计其态势，依据规则准确地预测可能发生的威胁、隐患，防患于未然，避免造成不良的后果或损失，而作出的态势评估，都迫切需要保证信任度，降低虚警率，避免或尽量减少误判/谎报现象的发生。

(2) 当意外的异常现象发生时，都迫切需要在其发生的第一时间内依据规则准确判断到底发生了何种异常现象，依据规则去可靠地处理异常现象，以提高对异常现象的有效检测与诊断。

由于规则可以是面向对象/集合/元组的，也可以是面向属性的，两者的层次高低不同，而我们研究规则的目的在于要基于规则进行推理，所以为满足信息融合的

需要，本书只研究面向属性的分类规则或决策规则。对此，目前大量的研究集中在知识发现以及机器学习等方法上，例如：

C4.5 算法、K. Micheline 提出的 MedGen 算法等。C4.5 算法在继承 ID3 优点的基础上进行了改进，能处理连续值类型的属性，它还能对属性的取值集合进行等价类划分，划分在同一类的属性值在属性值判断时将走到同一分支上。其思想简单使 C 4.5 在归纳学习中的地位更加显著。MedGen 算法在利用 C4.5 算法思想的基础上，采用了面向属性规约的方法对数据集进行预处理，以达到对数据集的“水平压缩”、“垂直压缩”，然后，建立决策树，其优点在于在一定程度上简化了决策树的建立过程。本质上，它们都是自顶向下的机器学习算法，即通过一组训练数据的学习，构造出决策树形式的知识表示，在决策树的内部结点进行属性值的比较并根据不同的属性值判断从该结点向下的分支，在决策树的叶结点得出结论，所有从根到叶结点的路径都对应着一条规则。基于决策树学习的算法的共同优点是在学习过程中不需要使用者了解很多的背景知识，但是，目前该类方法共同的不足之处有：

(1) 组成决策树的属性和属性值冗余性过大。

(2) 分而治之的处理策略存在较大的冗余性而产生了冲突现象。

(1)、(2)两个方面的不足直接导致了从决策树上提取的规则信任度较低。

不少学者基于粗糙集理论中的容差关系(对象集合 X 之上)的一个二元关系 R 称为一个容差关系(tolerance relation)，当且仅当关系 R 满足自反性和对称性。容差关系与等价关系的根本区别在于：① 允许对象之间可接受的差异；② 可以不具有传递性；③ 等价关系是特殊的容差关系(差异为 0)，它与概念格的泛化和特化存在的对应关系给出了将粗糙集理论和概念格的层次结构相结合并通过建立数据间泛化和特化关系来提取规则的方法。该类方法的优点是在一定程度上有助于提高格上规则挖掘的效率，是目前较有效的算法。但通过分类规则的提取结果不难发现还存在如下的问题：

第一，格上提取分类规则受到条件属性数量和个体对象数量的制约，当条件属性数量和个体对象数量较大时，如果不采用约束策略约简冗余信息，则提取的规则信任度较低。

第二，格上提取规则的过程存在较大的冗余性而产生的冲突现象降低了规则的信任度。

综上所述，针对同一问题对象的这两类方法，都存在因冗余性过大产生的冲突现象而导致提取的规则信任度较低的缺陷。

分析冲突现象产生的原因，可将冗余性的表现分为如下两种情况：

(1) 信息冗余。这是一种常见的冗余现象，存在的形式很多，如与问题对象无

此为试读，需要完整PDF请访问：www.ertongbook.com

关的原信息、产生的等效规则过多等。产生的原因也很多，如没有预处理或处理不彻底，相互包含现象等。

(2) 处理过程冗余。这是一种隐含式的冗余现象，与具体的处理结构和机制有关，例如，格结构遍历过程、构造树的方式等。处理过程冗余直接导致提取的规则存在不相容性、矛盾性等冲突现象。

为克服上述过大的冗余性产生的冲突现象而导致提取的规则信任度较低的缺陷，针对同一问题对象，本书提出一种新的格上和树中规则挖掘方法，拟将在动态信息融合框架下基于粗糙集(rough set)理论背景展开，具体研究内容如下：

(1) 说明面向属性的规则的含义、性质及规则挖掘的相关技术。

(2) 阐述采用什么样的机制来提取规则，这包括研究采用何种处理结构、规则挖掘的可能实现途径、提取后如何评估以及验证规则的可用性等。本书是在动态信息融合框架下研究规则的提取问题，因此，提出的机制应兼顾提取能力和动态信息融合必备的功能与目标。我们拟将核心提取机制分为三个上下文关联的部分：决策表的预处理组织层、提取更新层、简化精练层。强调该机制具有约简冗余的能力。

(3) 研究具体的规则挖掘方法。针对现有提取方法(如 C4.5 类的决策树方法、模糊近似方法、基于粗糙集理论的一些方法等)共同存在的不足，基于对不确定性的考虑，可基于粗糙集理论背景研究两种具有约简冗余能力的规则挖掘方法，即分明关系约束的格上规则挖掘法和基于包含度的决策树中规则挖掘方法，这两种方法均能提高规则的信任度，同时自身的结构特点使得它们比较适合动态信息应用环境。研究的内容涉及方法的基本处理过程、具体实现、性能分析比较等。

为保证对所研究的规则挖掘方法检验评估的真实性、合理性，可采用经典 D-S 证据理论进行融合验证。如图 1.2 示意了规则的提取与多元证据信息融合之间的关系，即利用粗糙集理论或模糊集理论等为背景提取带有信任测度的规则，基于规则推理后得到所关心的问题对象的状态判断，这些状态判断就是带有信任测度的各类证据，融合中心利用证据理论对多元证据信息进行融合，根据融合的结果估计问题对象的态势。

(4) 以水电厂线路工况为案例在仿真应用实践中验证本书所建议的机制和方法的有效性。

总之，在多种智能控制领域目前亟待解决的重要课题之一——用于决策的信任度高的规则挖掘问题框架下，需要研究如下两个承上启下的关键问题：规则挖掘机制问题、规则挖掘的具体方法问题。

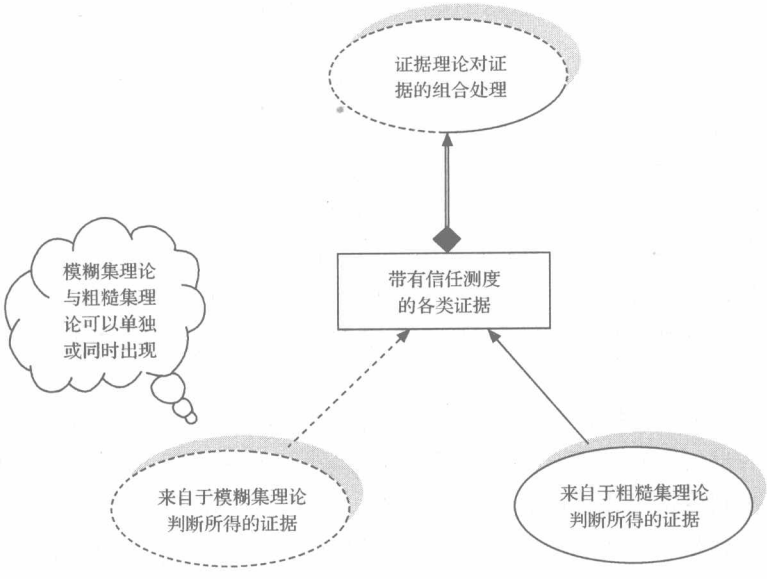

图 1.2 规则挖掘与多元证据信息融合之间的关系

1.3 规则挖掘技术的应用领域

规则挖掘技术的应用领域很多，如电子商务、机器人、交通管制、军事应用等。

动态多源异类信息融合中，规则挖掘技术问题显得十分重要。

按信息融合的目的不同可划分为两大类：目标状态信息融合和目标特性信息融合。这两类融合可以各自单独进行，也可以交叉进行。由于目标状态是动态变化的，而目标特性是相对稳定的，动态多源信息融合主要针对前者。

目标状态信息融合主要利用目标的功能、性能、变化情况等属性信息按时间的变化动态地进行，它适合动态运动/运行目标的跟踪、预测和估计。融合系统首先对传感器数据进行预处理以完成数据对准，在目标属性特征抽取的基础上再主要实现参数相关和状态特征向量估计或预测。目标特性信息融合主要利用目标的结构、形状等静态属性信息按空间进行，它适合静态或动态目标的识别。融合系统在融合前必须先对特征进行相关处理，把特征向量分成有意义的组合，如声波、图像融合等。

按信息融合的对象不同可分为：区化信息和类化信息。

区化信息即桶信息或块信息，通过区间的划分或范围的圈定而形成。范围的圈定如多源信息的聚类和合并，区间的分划如电流[0A，5000A]——[0，5]，[5，3000]，[3000，5000]——低，中，高。

通常情况下对连续数值信息的融合都要有区化这一步骤，连续值的区化具有如下意义：

（1）融合时存在阈值选择问题，由于连续值的取值任意性和数目的无限性，很难碰到刚好等于阈值的例子，那么分类处于临界点附近时就会错判，通常信息融合领域大部分问题具有信息残缺性的特点，这使得原信息级融合不能解决该类问题或效果不好，而区化后的特征级融合可以回避这一类问题，能够在示例信息较少的情况下获得较完备的先验知识，加快融合机制稳定的进程。

（2）提高信息有序程度，区化后的信息会变得更加有规律，但是这种有序有时是以丢失信息为代价的，单个区过大也会丧失这种有序性，因此要把握合适的区化程度。

（3）借助于人类融合方式的人工融合机制一般都要经过区化，因为人类对于原数据级的记忆和直接融合能力并不高，但却擅长抓住特征，如模糊方法的使用可以使融合机制更接近人类融合的工作过程。同时减少信息量，降低处理的复杂度。

类化信息是对原始采集的信息（包括区化后的数据和概念型信息）进行概念类化处理后形成代表一定范围的概念类，这是一个信息抽象提升的过程。它为融合算法提供了选择融合特征的空间。

例如，线路电压：$U=0\rightarrow U=$很低$\rightarrow U=$低，线路电流：$I=5000\rightarrow I=$太高$\rightarrow I=$高。

类化是信息融合的重要预处理方式之一，类化对于信息融合具有重要意义：

（1）类化是提取出了实例中的共同特征属性。

（2）由所有单体变为若干个集合，不同的特征级算法对特征的描述要求程度不同，而概念本身也根据抽象程度的不同分为若干类层。

（3）减小信息量，降低问题的复杂度。

理论上实现信息融合的常用基本方法如下：

（1）人工智能（AI）方法。专家系统（expert system）作为人工智能的一个应用分支，在信息融合的推理判决中得到了广泛的应用。其优点是可以模拟专家的经验知识、决策及推理过程，并用知识工程学构造其模型，产生相应的规则库，为高层分析或管理人员提供决策与施动的依据。如何产生规则库或者说如何提取规则是专家系统要研究的焦点问题之一，这是一个知识发现的问题。机器学习（machine learning）是一种重要的 AI 方法。

（2）模糊理论方法。在信息融合的不确定性推理中，模糊理论经常与其他理论联合使用。

（3）统计和推理方法。典型的如贝叶斯推理、D-S 证据理论等。

（4）聚类分析方法。它根据事先给定的相似标准，把观测值/样本分为类，按类分析，找出类内和类间的联系或规律性。

(5) 卡尔曼滤波法和估计(estimation)理论。卡尔曼滤波法主要用于融合底层的实时动态多传感器冗余数据。估计理论是数据融合与跟踪的基本理论,主要有经典估计理论(如最大似然估计、最小二乘法、贝叶斯估计等)和最优估计理论方法。

用于决策的信任度高的规则挖掘技术问题是动态多源异类信息融合领域目前亟待解决的重要课题之一。而现有的规则挖掘方法,大都存在因冗余性过大产生的冲突现象而导致提取的规则信任度不高的缺陷。该问题涉及的研究内容十分广泛且复杂,本书只对其中的几个关键问题进行研究:

(1) 规则挖掘机制;

(2) 规则挖掘方法;

(3) 规则挖掘过程中的分类等关键技术;

(4) 规则挖掘技术的应用案例。

此外的一些问题将后续进一步探讨,例如:

(1) 挖掘过程中某些阈值的优化问题。目前主要采用了领域先验知识的办法,这不可避免地增加了不确定性因素,后续准备加入研究其他的机器学习法等,通过训练,调整优化阈值。

(2) 复杂环境下多种方法的综合互补运用问题。对于多平台、多目标的复杂融合应用场景,由于同时在应用多种方法综合进行处理,保证多种方法间不相互干扰而协同互补地工作的问题也是需要研究的。对此,后续准备嵌入 CSCW 方法等对工作流进行协调。

(3) 机制与方法的通用性验证问题等。

1.4 规则挖掘技术的研究进展及内容

时间序列数据存在于社会的各个领域,如科学研究记录:天文观测、气象图像等;病历记录:患者每次看病的病情记录以及心电图等扫描仪器的数据记录等;金融和商业交易记录:如股市每天的交易价格及交易量、超级市场每种商品的销售情况等。时间序列几乎无处不在。随着科学技术的不断发展,计算机以及存储设备的存储容量日益增大,时间序列数据库也越来越大,对于时间序列的规则挖掘的研究也显得越发重要。

相对于规则挖掘较成熟的部分而言(如关系数据库中关联规则和分类规则的挖掘等),针对于时间序列规则挖掘的研究是数据挖掘研究领域中的较新的一个分支,目前国际上对于时间序列的规则挖掘的研究逐步成为一个新的热点,但国内在这方面的研究文献尚不多见,有一些学者曾经从理论框架的角度对时态规则挖掘做过介绍和分析。

时间序列的规则挖掘技术早在20世纪90年代已经有人提出，主要是对时间序列的相似性搜索，此后在此基础上又出现了时间序列数据库中子序列匹配和整体匹配的研究。在这些相似性计算过程中都存在时间序列数据量过大的问题，这将引起搜索效率的大大降低，因此时间序列的有效描述将是提高搜索效率的方法之一，由Pavlidis最早提出分段线性化描述是用直线段来近似拟合原始时间序列的形状，这样用直线段来代替原始数据将大大减少数据量，此后Keogh又作了进一步的研究，但是有些文献中提到的分段线性化方法计算量较大，需要多次迭代，这一缺点在数据量大时尤为明显，而且其中采用的相似性测量方法对于时间轴按比例缩放的情况是敏感的，然而这一现象在实际时间序列中是经常出现的，因此解决这一问题也是相当重要的。

在相似性的基础上，人们又发展出了许多模式识别的方法，如通过相似性计算对时间序列进行分类、聚类等。关于时间序列的规则挖掘技术除了相似性的研究外，还有其他的一些更接近于人类思维方式的研究方法即模式挖掘方法，如关联规则的抽取，可以得到形如“如果某一天Microsoft上涨而且Intel下降，则IBM第二天上涨”的规则，而有些文献提出可以从时间序列中抽取分类规则，但是这种方法有一个前提条件，即是必须对时间序列进行预处理，从中预先抽取静态模式，然后从模式中提取对时间序列影响较大的特征属性，对时间序列进行趋势预测。这种分类规则的提取对于科学观测数据的分析以及金融经济的预测将会提供有效的帮助。但是有些文献中采用的规则挖掘方法得到的预测精度比较低，有待于研究更有效的、泛化能力更强的分类规则挖掘方法。

前馈神经网络作为规则挖掘的分类工具，已有很多论文发表，但是前馈神经网络固有的缺点（如过拟合、分类规则的知识表示无法理解等）使得人们提出了各种各样的改进模型以便提高泛化性能或从训练好的神经网络中提取分类规则。但现有算法存在各种各样的局限性，有待进一步的改进以提高适用性。同时可以将神经网络和其他不同的分类技术如粗糙集理论、模糊逻辑、机器学习等规则生成技术进行信息融合，互相取长补短，提高分类精度。

属性分布随时间变化的时变数据也存在于很多领域，如半导体制造过程是经常在工程师的指导下做出改变的，以便快速地达到较高的生产率，从上个月数据中抽取的知识不可能正确地预测本月的状态，而普通的分类技术将不再适用于时变数据的挖掘，因此提出一种针对于时变数据的分类技术显然是很必要的，但迄今研究的人并不多，本书的研究将为时变数据库的分类技术提供一种有力的工具。

时间序列的规则挖掘技术自20世纪90年代中期以来有了快速的发展。由最初的相似性的分析到目前的人工智能的多学科交叉研究，时间序列的规则挖掘技术已经有了多个研究方向。

相似性的研究是时间序列的一个最基本的而且比较困难的问题。所谓相似

性，简单地说是指测定两个给定的时间序列是否具有相似的行为曲线，这个问题之所以困难是因为时间序列往往是来自于实际，因此对于相似性的测量要求并不是完全严密的，而且时间序列数据库来自于各个领域，测量标准也不尽相同。规则挖掘技术的不断发展，产生了各种挖掘技术，因此这些技术在相似性的基础上也不断地被应用于时间序列数据库，如对时间序列进行聚类、分类，产生关联规则等。

时间序列研究的另一个基础问题是时间序列的索引(index)问题，即给定一个时间序列集 Q，索引技术是指能够在 Q 中尽快地找到与给定时间序列 q 最相似的序列子集 q_1，当然索引技术也需要相似性的计算。

无论是相似性研究还是索引技术，时间序列的有效的描述(representation)仍然是提高计算效率的一个关键途径。

关于时间序列的规则挖掘技术除了相似性的研究外，主要还有模式挖掘的研究，其中主要包括时态模式挖掘和趋势预测。时态模式挖掘的一个主要技术是关联规则的挖掘。趋势预测采用的主要是分类规则的挖掘技术，即 M. Last 提出的首先对时间序列进行预处理，然后从中抽取关键的预测属性(predicting attributes)，这些属性对时间序列的发展趋势影响较大，将其组成属性集，这些预测属性表征了时间序列的某种特性，这种特性与时间没有关系，因此可以采用普通的静态的规则挖掘工具对时间序列进行行为趋势的分类预测。

本书针对上文提到的几个关键问题进行了如下研究：

(1) 研究了具有冗余约简能力的规则挖掘机制。该机制同时兼顾了规则挖掘能力和动态信息融合应具备的功能，强调通过融合验证评估提取规则的有效性以及提取过程是增量式的、上下文信息是关联的、信息间是协同互补的。

(2) 研究了“分明关系约束的粗糙格上规则挖掘方法”。改进了增量式构建格空间的算法，设计了利用分明关系约束提取信任度高的规则的算法，克服了传统同类方法存在的“无约束状态下存在的遍历结点冗余产生的冲突现象而导致提取的规则的信任度不高”的缺陷。

(3) 研究了“基于包含度的决策树中规则挖掘方法”。基于包含度改进了选择作为决策树的子结点的新属性的方式，通过剪枝调整了决策树的深度和结构，设计的基于包含度的决策树构建算法，克服了现有同类方法存在的“树结点冗余产生的冲突现象而导致提取的规则的信任度不高”的缺陷。

(4) 基于时间序列相似性的研究，研究了与时间序列结构自适应的分段线性化表示方法，该方法可以自动地产生拟合时间序列的直线段的段数 K，并且提出了基于这种表示方法的相似性测量公式。其优点在于：不必事先提供分段线性化后直线段的段数 K；能够滤去实际时间序列数据中的噪声；大大压缩了相似性计算的计算量；相似性测量公式能够适应多种时间序列的变形。

(5) 基于趋势预测的研究，研究了将正则神经网络应用于时间序列行为趋势

此为试读，需要完整PDF请访问：www.ertongbook.com