干预规则挖掘的概念、任务与研究进展

段 磊[1)] 唐常杰[1)] 杨 宁[1)] 左 劼[1)] 王 悦[2)] 郑皎凌[3)] 徐开阔[4)]

[1)](四川大学计算机学院 成都 610065)

[2)](北京大学信息科学技术学院 北京 100871)

[3)](成都信息工程学院软件工程系 成都 610225)

[4)](成都信息工程学院计算机系 成都 610225)

摘 要 干预规则挖掘是近年从干预实践中提出的新型数据挖掘任务,旨在利用数据挖掘技术探测干预事件,发现最佳干预时机和力度,提供促进事物向期待状态转化的决策支持.文中以四年的研究实践为背景,介绍干预规则挖掘的研究沿革和现状,给出了干预规则挖掘的任务分类.从三个角度,即干预效果预测、干预方法发现和未知干预探测三方面,介绍干预规则挖掘的研究问题、困难和成果.展望了干预规则挖掘未来研究方向.

关键词 数据挖掘;干预规则;流数据;不确定数据;时间序列

中图法分类号 TP311 **DOI 号**: 10.3724/SP.J.1016.2011.01831

Concepts, Tasks and Research Advances of Intervention Rule Mining

DUAN Lei[1)] TANG Chang-Jie[1)] YANG Ning[1)] ZUO Jie[1)] WANG Yue[2)]

ZHENG Jiao-Ling[3)] XU Kai-Kuo[4)]

[1)](*School of Computer Science, Sichuan University, Chengdu* 610065)

[2)](*School of Electronic Engineering and Computer Science, Peking University, Beijing* 100871)

[3)](*Department of Software Engineering, Chengdu University of Information Technology, Chengdu* 610225)

[4)](*Department of Computer Science, Chengdu University of Information Technology, Chengdu* 610225)

Abstract Intervention rule mining is an emerging data mining task, which is derived from the practice of intervention application. It aims at applying data mining techniques on detecting intervention events, discovering the best intervention time and intensity, and decision support for converting objects from undesirable state to desirable state. This paper introduces the research background, as well as the major related advances on intervention rule mining based on the four-year practice, and defines the task classification. Moreover, this paper surveys the research issues, difficulties and achievements in three aspects, i.e. intervention effect prediction, intervention method discovery, and unknown intervention event detection. Finally, this paper discusses the future work of intervention rule mining.

Keywords data mining; intervention rule; data stream; uncertain data; time series

1 引 言

1.1 干预规则挖掘的提出

客观世界不以人的主观意志转移,客观事物的运行状态包含人们期望的和非期望的两类状态.例如:新生婴儿有健康和不健康的,肿瘤有良性和恶性的,企业有盈利和亏损的,等等.人们不能改变客观规律,但可以在发现规律的基础上,促进事物向期望的方向转变,即循律促变.例如,对出生缺陷进行干

收稿日期:2011-08-12;最终修改稿收到日期:2011-09-15.本课题得到国家自然科学基金(60773169)、"十一五"国家科技支撑计划(2006BAI05A01)、高等学校博士学科点专项科研基金(20100181120029)、四川大学青年教师科研启动基金(2009SCU11030)资助.**段 磊**,男,1981 年生,博士,讲师,中国计算机学会(CCF)会员,主要研究方向为数据挖掘、进化计算.**唐常杰**(通信作者),男,1946 年生,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为数据库、数据挖掘.E-mail:cjtang@scu.edu.cn.**杨 宁**,男,1974 年生,博士,讲师,主要研究方向为数据挖掘.**左 劼**,男,1977 年生,博士,讲师,主要研究方向为数据挖掘.**王 悦**,男,1981 年生,博士,研究方向为数据挖掘.**郑皎凌**,女,1981 年生,博士,讲师,主要研究方向为数据挖掘.**徐开阔**,男,1983 年生,博士,讲师,主要研究方向为数据挖掘.

预,提高健康婴儿出生率;希望将恶性肿瘤治愈为良性;希望扭亏转盈.

干预规则挖掘(Intervention Rule Mining)应运而生,它是新的数据挖掘任务,是循律促变的技术.干预规则挖掘是从出生缺陷干预、糖尿病干预、带约束的市场调控等实际问题中抽象出来的新基础性课题,其核心研究目的是在认识自然的基础上,利用数据挖掘技术发现和验证改造自然的方法和措施,发现并遵循事物发展的动力学规律,施加干预,使被干预对象向人们期望的方向发展,体现了循律促变的思想.

干预规则挖掘含 3 个要点:(1)挖掘指定对象的干预动力学行为规律,即干预规则;(2)发现指定对象在给定干预下的响应规律,分析最有效的干预时机和力度;(3)见微知著,从对象的微变分析,发现和预测外界干预因素.将分析结果应用于工程设计、科研实践、社会调控(金融体系、国家政策评价)等领域,为决策提供依据.

应用干预规则,决策者可预测某种决策,在何种条件,何等规模下的干预,可得到何等干预效益.如果说传统数据挖掘是认识世界,干预规则挖掘就是为人类改造世界做技术准备.

干预规则挖掘涉及了分类、预测、关联、多维数据分析、函数发现等多项数据挖掘技术.但干预规则挖掘并不是这些技术的简单叠加,而是在这一系列方法基础上的一类新任务.

1.2 干预规则挖掘的现实应用

干预规则挖掘具有广泛的应用领域和较高的应用效益.这里列举干预规则挖掘在卫生、经济、社会领域中的三项典型应用.

出生缺陷干预. 我国是出生缺陷高发国家,对出生缺陷进行干预是提高出生人口素质,防止、减少出生缺陷危害的重要举措.20 多年出生缺陷监测数据表明,全国各地区在适宜干预技术的研究、应用和推广方面发展不均衡,且出生缺陷的发生存在病种、地区、人群间的差异,对特异性和普适性出生缺陷干预技术的需求并存.因此,利用我国已建立的出生缺陷监测数据库和完善的大型流行病学调查现场,更准确地摸清全国出生缺陷发生状况和变化规律,了解不同地区的出生缺陷干预措施实施情况,掌握影响干预效果的医学和社会学因素,评价干预效果,是开展出生缺陷干预并获取最大干预效果的必要前提.我国自 1999 年启动国家新生儿"出生缺陷干预工程"提倡以增补叶酸预防神经管畸形.此外,对于具体的干预措施,还应考虑如何实施最合理,效果最好.例如:增补叶酸需考虑在什么时机补?补多少?补多久?

房价波动及其干预. 房价始终是一个受民众关注的问题,房价的高低受到包括:竣工房屋造价、当地人均可支配收入、地方税收行政费用等在内的多方面因素影响.政策性的干预措施是保障房地产健康发展、抑制房价上涨过快的必要手段,可采用何种手段作为干预措施,实施的力度、代价与实施效果的评价都是尚待解决的难题.

贫富差距及其干预. 中国人民大学的一项研究表明中国富裕家庭与贫困家庭的收入相差达几十倍;城镇化进程加快的同时,城镇贫富差距问题比农村还严重.因此,制定并实施缩小贫富差距的干预措施是建设和谐社会、实现共同富裕的当务之急,国家制定了一系列干预措施,包括新型财产税税收体系,消费税改革,调节个人所得税起征点,适时开征物业税,完善社会保障制度,加强城乡扶贫工作,消除不平等竞争等.每项措施的实施方式、力度、代价、效果评价是决策部门面临的难题.

1.3 干预规则挖掘作为数据挖掘任务的背景

鉴于干预规则在现实世界生产活动的广泛应用,四川大学计算机学院数据库与知识工程研究所自 2007 年开始同中国出生缺陷监测中心合作对我国连续 20 年出生缺陷监测数据进行数据分析工作,并以此为基础提出把干预规则挖掘作为一项新的数据挖掘任务,用现代数据挖掘方法和工具进行研究.在国家自然科学基金、国家"十一五"科技支撑计划、教育部博士点基金的支持下取得了一系列科研成果:探索了朴素干预规则和数值型干预规则挖掘算法、数据流干预分析模型、个体疾病状态干预、群体疾病状态干预,基于数据流的未知干预发现技术,并行事件序列干预规则挖掘等,并开发了出生缺陷数据挖掘系统 HealthyBaby 1.0,正在实践中不断改进[1-15].

本文首先概述干预规则挖掘的基本概念,介绍了同干预规则挖掘相关的分析方法,然后指出干预规则挖掘的基本流程及任务分类,接着围绕干预规则挖掘任务的分类讲述了干预规则挖掘的主要研究问题及相关进展,最后总结全文并展望未来的工作.

2 干预规则挖掘相关其它干预分析

2.1 统计学中的干预分析

从传统统计学角度研究干预的分析模型最早由 Box 等人在文献[16]提出,并迅速被应用于描绘突发事件或政策变化对经济或环境所产生的影响.对

突发事件或政策干预造成的影响作定量分析是十分重要的.例如,我国 1978 年对农业实行经济体制改革后,农业产量显著上升,使绝大多数地区解决温饱,迈向小康生活,这是政策干预的结果.

ARIMA 模型从序列自相关的角度揭示时间序列的发展规律.通常情况下,序列值之间存在一定的相关关系,而且这种相关关系具有一定统计规律.ARIMA 模型分析的要点在于寻找这种规律,并拟合适当的数学模型来描述这种规律,进而利用这个拟合模型来预测序列未来的走势.因此,大多数统计学研究采用 ARIMA 模型进行干预分析.文献[16]将 ARIMA 模型应用到经济、环境问题的干预分析中,文献[17]对 ARIMA 模型进行扩展,并应用到股票及工业统计数据分析中.此外,其它统计模型如:指数分布、泊松分布也被用于干预分析.例如:通过扩展指数分布检查邮件流中的突发事件[18];根据泊松分布假设提出自适应方法检测车流量[19].

2.2　与干预分析相关的数据挖掘技术

What-if 分析根据已知历史数据,用"假设采用某措施,则会有某期待结果"的思维方式,分析或发现未知结果的可能性[20].What-if 分析可以在数据库、数据仓库中的关系表和 OLAP 数据库中的多维数据上进行[21].作为一个重要的决策支持辅助功能,what-if 得到了广泛的应用.通过 what-if 可以解决两类决策支持问题:(1)通过拟合历史数据,预测决策方案执行情况;(2)在真实的历史数据上,评估未执行的决策方案.What-if 分析同干预规则挖掘都以决策支持为目的,但在具体任务和实现方法上存有区别.有关在 OLAP 系统中进行 what-if 查询处理和支持 what-if 分析的 OLAP 系统的研究也得到了广泛的关注[20-23].文献[20]提出了维度层次结构下的 what-if 类查询以及执行 what-if 查询的系列代数操作;文献[21]介绍了若干 what-if 分析的实现方法;文献[22]针对主存 OLAP 系统,设计了基于内存记录指针的方法提高 what-if 数据视图的合并性能以及 what-if 分析的代价模型.文献[23]分析了 what-if 分析中多版本数据预处理和数据立方增量计算的问题.What-if 查询和分析需要基于历史数据建立多场景假设,delta 表可以用来记录复杂、多版本的假设更新,文献[24]针对传统 delta 表合并效率低的不足,基于集合操作对其进行了优化.此外,文献[25-26]研究了在商业智能领域建立 what-if 分析模型的方法,提出了基于扩展 UML 进行概念表述的建模方法.

文献[27-29]考虑了可行性挖掘(actionability mining)的问题.文献[27]从包含若干属性和实施效力的历史数据中,发现一组能够对目标个体提高实施效力的行为模式.文献[28]考虑了客户关系管理上的一个问题,将"发送信件"或"给家里打电话"看作是行为,研究维护、提升客户关系的方法.文献[29]基于决策树方法设计一个贪心算法,以寻求能够最大化利润的一组行为.文献[30]提出利用显露模式[31]发现潜在可能用于基因疗法的基因,通过改变这些基因的基因表达值治疗肿瘤.文献[32-33]研究了从称为旧数据集和新数据集的两个数据库中,挖掘分类模型的显著变化的方法.文献[34]提出了从高端和低端实例中挖掘用户偏好的方法.

3　干预规则挖掘任务及亚复杂系统建立

3.1　干预规则挖掘任务分类

图 1 示例了实践中干预规则挖掘、评价及实施的全流程.对经过预处理的历史数据建立亚复杂系统,应用干预规则挖掘算法发现满足干预需求的若干干预规则.由人工或计算机对发现的干预规则进行评价,筛选出的若干干预规则组成具体的干预措施.预测干预措施的干预效果能辅助决策干预措施的实施方式.干预措施实际执行后可对其干预效果进行验证.

图 1 表明,干预规则挖掘并不代替人进行干预决策,而是利用数据挖掘方法和技术,通过为制定干预措施提供候选干预规则达到决策支持的目的.直观地,现实的干预应用通常涉及诸多因素是个复杂系统,挖掘可行、有效的干预规则需要引入领域知识和必要的人工评价策略.

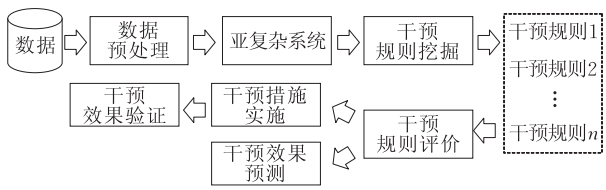

图 1　干预规则挖掘及评价、实施的基本流程

为简明地表达干预规则挖掘,本文给出必要的符号和术语,避免复杂的形式化描述,并通过例子说明思路.设被干预对象记为 o,其状态记为 S_p,干预措施为 F,干预措施的实施方法为 ε(如强度、频率等),预期干预效果为 S_e(干预实施后 o 的预期状态),实际干预效果为 S_i(干预实施后 o 的实际状态).干预行为表达为

$$Inv(o,F):S_p \xrightarrow{F(\varepsilon)} S_i.$$

根据干预规则挖掘的应用需求，我们将干预规则挖掘分为三类基本任务，下面举例说明.

干预效果预测. 已知当前状态 S_p、干预措施 F，预测干预 F 实施后，干预效果 S_i.

例 1. 已知某地区出生婴儿神经管缺陷的发病率，预测对产妇补充一定量叶酸(降低出生婴儿神经管缺陷的一种措施)后，该地区出生婴儿患神经管缺陷的发病率.

该类任务的另一种表达形式是已知当前状态 S_p、干预措施 F，预测干预 F 实施后有多大概率达到预期干预效果 S_e.

干预方法发现. 已知当前状态 S_p，预期干预效果 S_e，挖掘能到达干预目的的干预措施 F.

例 2. 已知对产妇补充叶酸能降低出生婴儿神经管缺陷的发病率，要使得某地区出生婴儿神经管缺陷发病率降低 p（预期干预效果），对每个产妇应该补充多大剂量的叶酸？

该类任务的另一种表达形式是已知当前状态 S_p、干预措施 F，求达到预期干预效果 S_e，干预措施的实施参数 ε 的取值.

未知干预探测. 已知历史状态 S_p、当前状态 S_i，判断是否有干预发生，发生干预的形式是什么.

例 3. 已知持续监测某地区某类出生缺陷发病率的变化，判断是否有干预事件发生，并挖掘出所发生的干预事件.

该类任务的另一种表达形式是在若干干预措施中探测发挥干预效力的干预措施(干预事件)F.

上述三类任务描述了目前干预规则挖掘研究的主要内容. 新的研究内容，随干预领域的扩展、研究与实践的深入，将不断被提出. 我们在实践中，根据数据对象的性质，把干预规则挖掘任务分为

(1) 静态数据的干预规则挖掘；

(2) 流数据对象的干预规则挖掘；

(3) 不确定数据的干预规则挖掘.

根据对象分类基数，可将干预规则挖掘分为：

(1) 单个体干预规则挖掘. 例如：对某一特定产妇进行出生婴儿缺陷干预.

(2) 类个体干预规则挖掘. 例如：对某一地区所有产妇进行出生婴儿缺陷干预.

此外，根据干预属性和被干预属性的类型，干预规则可分为范畴型、连续型和混合型. 当干预属性为非数值型时，使用范畴型干预规则描述. 这里我们用出生缺陷干预为例进行说明，事实上干预规则挖掘可广泛应用于其它领域.

3.2 亚复杂系统

观察被干预对象的状态与时态，可看出干预的本质是通过外界作用改变事物的状态(属性)，进而改变事物状态序列发展的轨迹，使其朝着期望的方向发展. 改变的状态轨迹同原状态轨迹之间的差异即是干预的效果(效力). 复杂系统研究称满足下列 3 个特征的对象为复杂系统：(1) 系统行为形似随机而实非随机；(2) 系统行为由其内在规律决定；(3) 简单对象构成却有复杂行为表现. 同时，复杂系统具有自适应性、局部性、竞争性、突变性、非线性和不确定性.

对现实世界事物进行干预，干预作用会利用客观的动力学规律影响事物发展趋势. 干预过程中由于受到各种潜在因素的影响，可能会表现出一定的随机性. 实践中的干预对象是复杂系统. 复杂系统常具有高度突变性、非线性和不确定性，目前技术较难在真正意义上有效解决复杂系统领域的问题.

对复杂系统进行合理简化，是人们一贯的方法. 通过特征提取、忽略次要因素、降维等处理，得到一个相对简单、剥离混沌性质且目前科学能力和现有技术可能做出工程性解决方案的系统，称为亚复杂系统. 亚复杂系统的建立为描述人工干预下亚复杂系统动力学行为，揭示干预普适规律，并提出可工程化的解决方案提供了基础. 我们称应用于干预规则挖掘的亚复杂系统为干预分析亚复杂系统.

为干预分析建立的亚复杂系统是为了挖掘隐藏于其中的干预规则，并对相关干预行为进行动力学分析. 建立干预分析亚复杂系统的要点包括：

(1) 干预相关度分析. 干预分析模型中包含了众多相关属性，相关度大的干预属性适宜进行并发干预研究. 相关度低的干预属性可分离到不同干预分析模型中，降低问题难度.

(2) 亚复杂系统属性选取. 在复杂系统基础上建立干预分析亚复杂系统的基本步骤包括：① 去除冗余属性；② 去除干预相关度低的属性.

(3) 确定干预分型. 干预规则分为多种类型，可有多种表现形式. 如何进行干预分型，采用合适的干预类型描述干预规则，需结合应用需求和领域知识确定. 干预分型体现了分而治之策略，有助于降低分析难度，提高干预规则准确度. 同时，也适宜于不同方法进行干预分析和干预规则挖掘.

(4) 干预规则评价. 干预规则评价是干预分析亚复杂系统的重要组成部分，度量了干预规则挖掘结果和预期发现目标之间的差异，体现了干预规则挖掘的准确度和有效性.

将复杂的现实世界干预问题，抽象为可用计算机分析处理的亚复杂系统，是进行干预规则挖掘的基础.每一项干预分析问题都有其独立的干预分析模型.干预分析亚复杂系统也是干预规则评价和解释的基础.有关从复杂系统提取干预分析亚复杂系统的研究，请参考文献[1,6,12].

4　干预效果预测研究

4.1　朴素干预规则

考虑一个典型的营销决策问题："以历史数据为依据，如果增加广告投入(x)，产品的市场占有率(y)是否会提高，提高多少?"这正是干预效果预测拟解决的问题，即通过分析干预属性和干预目标之间的因果关系，预测干预属性发生变化后，干预目标相应的变化.由于关联规则能反映项集之间的关联信息，文献[3]提出了基于关联规则的朴素干预规则模型，描述干预属性和干预目标之间的变化关系.朴素干预规则同关联规则的区别在于关联规则中的项是静态不变的，而朴素干预规则需考虑干预改变属性取值的情况.

朴素干预规则源于这样一个思想："朴素干预规则＝关联规则＋增量分析(或微分扰动)"，即在关联分析基础上考虑变化的因素，形式化描述如下：可进行干预操作的属性称为干预操作集，记为 $A=\{A_1, A_2, \cdots, A_m\}$，反映干预目标的属性称为干预效果集，记为 $E=\{E_1, E_2, \cdots, E_n\}$，$A \cap E=\varnothing$.干预操作集中项集记为 $Dom(A)$，$Dom(A)=Dom(A_1) \cup Dom(A_2) \cup \cdots \cup Dom(A_m)$，干预效果集中项集记为 $Dom(E)$，$Dom(E)=Dom(E_1) \cup Dom(E_2) \cup \cdots \cup Dom(E_n)$.朴素干预规则是满足下列条件的表达式：

(1)存在频繁项集 $IS \subseteq Dom(A)$，IS 中的属性为 $Att(IS)=\{A'_1, A'_2, \cdots, A'_d\}$，存在 $e \in Dom(E)$，满足 r'：$s_1 \wedge s_2 \wedge \cdots \wedge s_d \Rightarrow e$ 是强关联规则，其中 $s_k \in Dom(A'_k)$，$1 \leq k \leq d$，支持度记为 $Sup(r')$，置信度记为 $Conf(r')$.

(2)对 $Att(IS)$ 中的每个属性，存在 $t_k, s_k \in Dom(A'_k)$，$1 \leq k \leq d$，且 $t_k \neq s_k$，满足 r''：$t_1 \wedge t_2 \wedge \cdots \wedge t_d \Rightarrow e$，支持度记为 $Sup(r'')$，置信度记为 $Conf(r'')$.

(3)设项 I_k 表示 $s_k \rightarrow t_k$，$1 \leq k \leq d$，项 I_e 表示 $e \rightarrow \neg e$，则 r：$I_1 \wedge I_2 \wedge \cdots \wedge I_d \Rightarrow I_e$.

(4)规则 r 的支持度 $Sup(r)$、变化度 $Delta(r)$ 和置信度 $Conf(r)$ 分别定义为

$Sup(r)=Sup(r')$；

$Delta(r)=Conf(r')-Conf(r'')$，$Delta(r)>0$；

$Conf(r)=Delta(r)/Conf(r')$.

上述规则中，$Sup(r)$ 体现了干预影响的范围；干预操作集中的属性值发生改变后(即干预)，所产生扰动的方向和幅度用 $Delta(r)$ 描述，$Conf(r)$ 是变化量在原始状态下的比值，体现了变化的准确率.

朴素干预挖掘算法的主要步骤包括：

(1)指定干预目标属性，挖掘包含干预目标属性的频繁项集.

(2)对频繁项集中包含的各属性，计算同属性中不同项之间的关联规则，挖掘出变化量($Delta$)最大的朴素干预规则.

在提出朴素干预规则的基础上，我们团队应用朴素干预规则于 1986 年至 1987 年全国出生缺陷监测数据，并发现若干有意义的规则[3]，下面是实验结果中置信度最高的 2 条规则.

规则 1.　时间(1986～1987)|近亲结婚(父母→否)⇒缺陷儿(是→否)　$Sup=0.0240$，$Delta=0.3388$，$Conf=0.9906$.

规则 2.　时间(1986～1987)|先天患病(有→无)⇒缺陷儿(是→否)　$Sup=0.0005$，$Delta=0.4622$，$Conf=0.9307$.

4.2　基于拟合函数的干预效果预测

由于朴素干预规则基于关联规则分析建立，所以其分析对象的数据类型为离散型.数值型的分析对象需要事先进行离散化.为满足实际应用，我们团队提出了直接针对数值型数据的干预规则挖掘算法[1].数值型干预规则挖掘算法的要点包括：(1)将干预操作集和干预目标属性分别作为自变量和因变量；(2)判断自变量和因变量是否相关，对相关的属性集，挖掘拟合函数；(3)分析拟合函数的单调性，分区间计算支持度、变化度和置信度.

为了判断自变量属性和因变量属性是否相关，文献[1]将数据集分为两组，分别进行函数拟合，对得到的两个函数，比较它们的泰勒公式系数，如果差异大于阈值则认为考查的属性不相关，否则认为这两个函数"大致相同"，考查的属性相关.函数中拟合程度较高者作为数值型干预规则的函数.根据对函数求一阶导数的结果，划分函数的单调区间，即可得到各单调区的单调性(ASC 或 DESC).考虑到数值型干预规则挖掘中训练数据的特点，文献[1]采用基因表达式编程[35]做非线性主成分分析.

为了更好地描述干预规则，文献[1]定义区间的支持度(Sup)来反映区间的重要性；变化度($Delta$)为拟合函数的一阶导数，反映函数变化的趋势；置信度($Conf$)是主成分属性个数与所有属性个数的比

值.在1986年至1991年全国出生缺陷监测数据上,文献[1]发现如下两条规则.

规则3. 时间($a\in[1986,1991]$)⇒围产儿死亡率($-0.872a+102.330$) DESC, $Sup=1$, $Delta=-0.872$, $Conf=1.0$.

规则4. 围产儿死亡率(‰)($a\in[22.8,26.7]$)⇒死亡中缺陷率(%)($-1.456a^2+65.874a-697.730$) DESC, $Sup=1$, $Delta=-2.912a+65.874$, $Conf=1.0$.

在出生缺陷监测数据上的实验表明,数值型干预规则对分析新生儿统计信息有较大作用,它能挖掘出具有相关性的数值属性,单调性分析可为决策者提供决策支持.

4.3 不确定数据集上挖掘优化的概率干预策略

现实世界的观察数据,由于设备、方法、人为等原因不可避免地导致与真实数据出现一定程度偏差,因此数据带有不确定性.不确定数据管理是目前研究的热点,在不确定数据上分析评价干预效果同样具有不确定性,如何获取干预效果的不确定性是干预规则挖掘的新问题.文献[14]提出不确定数据集上挖掘优化干预策略的问题,采用"假设策略→真实历史数据→得出结论"的方式分析带有不确定性的历史数据,从中探索假设干预策略的可行度.即在不确定数据上,建立干预策略评价模型,给出概率性干预效果的评估和某种干预效果的概率,为决策提供量化支持.解决类似"若实施干预策略 s,可能对结果产生什么影响"的问题.

不确定数据上干预策略评测的两个问题[14]:(1)干预效果预测评价,对指定的干预措施预测其实施后的可能效果;(2)计算指定干预策略达到预期干预效果的概率.具体地,文献[14]采用谓词逻辑表述干预策略,获取该谓词在历史数据中的统计信息,并利用这些信息分析干预策略可能的效果.

干预策略强度是干预策略有效性的量化表现,反映在实际中为响应它的实例数.例如:事后评定一项出生缺陷干预策略的效果,可考察策略实施后,出生缺陷婴儿是否减少.困难在于,预测一项干预策略的执行效果,干预尚未实施,没有效果可评;解决的思路是,历史数据中,常常存有类似干预的扰动,检查其后的效果,相当于在历史上曾经有意无意地采用了该干预.这样得到的评价意见是以历史可重复原理为依据,以历史事实为准绳.

根据这个思路,文献[14]定义了干预强度及影响.

定义1. 设 U 是带有不确定性的实例数据.给定时间间隔 t 和干预策略 s,s 的长度为包含的不同谓词数:① 给定评价算子 e,若数据项 u 满足 e 记为 $e(u)=\text{true}$;② 干预策略 s 在 t 的强度为 t 内满足 s 的实例数与 t 内所有实例数的比值,记为 $\gamma_{s:t}$,$\gamma_{s:t}=|\{u\mid u_{\text{time\_interval}}=t,\ s(u)=\text{true},\ u\in U\}|/|U|$;③ 干预策略在 U 的影响为:时间段 t 内满足 e 的数据实例数与 t 内所有实例数的比值,记为 $\theta_{s:t}$,$\theta_{s:t}=|\{u\mid u_{\text{time\_interval}}=t,\ s(u)\cap e(u)=\text{true},u\in U\}|/|\{u\mid u_{\text{time\_interval}}=t,\ s(u)=\text{true},\ u\in U\}|$.

由于外界因素的存在,即使相同的干预在不同时间段的干预效果也存在差异.为描述这种差异,文献[14]提出 p 概率干预策略.

定义2(p 概率干预策略). 对不确定监测数据 U,若干预策略 s 在给定时间段集合 $T=\{t_1,t_2,\cdots,t_n\}$ 中以概率 p 使其干预影响值($\theta_{s:t_i}$)满足 $\theta_{s:t_i}\geq k$,则称 s 为限度 k 的 p 概率干预策略,记为 $p\text{-}s:k,p_r(s)=|\{\theta_{s:t_i}\mid\theta_{s:t_i}\geq k,\ 1\leq i\leq n\}|/n$.

p 概率干预策略可更精确地度量干预效果,可以知道在一系列时间段上达到某个指标的次数.

为满足海量数据处理需求,文献[14]还提出谓词统计树、维度排序等方法优化干预策略评价算法.

5 干预方法发现研究

5.1 疾病状态干预

基因疗法和药物干预[36-37]相关研究表明:通过外力更改病变组织的基因表达水平(gene expression level),恢复其正常水平,有望达到治疗疾病的目的.对此,文献[9]提出疾病状态干预,对给定的病例样本集(包含不良状态和良好状态)挖掘可能将疾病状态从不良转变为良好的干预方法,从而为实施疾病治疗提供决策支持.干预通过改变对象属性值的方法来实现,涉及到的属性及其改变的目标值称为状态转换项.例4与图2说明了疾病状态干预问题.

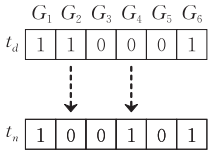

图2 疾病状态干预示意

例4. 设有6-基因($G_1,\cdots,G_6$)的病例样本,$t_d=(1,1,0,0,0,1)$是一不良状态样本,$t_n=(1,0,0,1,0,1)$是一良好状态的样本.其中,样本中的0、1分别代表基因表达水平为低或高.如图2所示,若将 t_d 中基因 G_2 和 G_4 的表达水平分别从高变低和从低变高,样本 t_d 将从不良状态转换为良好状态.因此 G_2、

G_4称为状态转换项，组成的集合称为状态转换集.

例 4 表明，属性的变化会引起样本状态的改变. 显然，实用性高的干预方案应包含尽量少的属性，并且包含属性的改变代价应尽量小. 文献[9]设计了状态转换项挖掘算法，包括 3 个要点：(1) 选择候选状态转换项，(2) 在候选状态转换项中进行比较，选择有效的状态转换项，(3) 排除干预效果已被状态转换集中状态转换项覆盖的候选状态转换项.

根据被干预对象是一个或一类病例，疾病状态干预分为个体疾病状态干预和群体疾病状态干预.

5.2　疾病状态转换项挖掘算法

5.2.1　个体疾病状态干预

对于给定的一个不良状态的病例样本 o_u，状态转换项挖掘算法要点如下：

(1) 确定待改变的属性及其改变的目标值. 对任意两个病例 o_1、o_2，差异度为不同属性值的个数；

(2) 确定 o_u 中需干预的属性. 选取 k 个与 o_u 最相似且处于良好状态的样本，通过对比找出待改变属性. 对每个待改变属性的目标值，参照良好状态的样本依条件概率选取最佳目标值；

(3) 候选状态转换项排序. 考虑：① 属性 A 在样本 o_u 中的值与不良状态的关联度如何？② 良好状态样本中有多少样本在属性 A 上取值与 o_u 不同？

为成功转换病例状态，希望被干预属性的当前值与不良状态强相关，目标值与良好状态强相关且与不良状态弱相关. 同时，为达到最佳干预效果，应尽量选取相关性小的状态干预项组成状态干预集. 文献[9]采用增量式的方法寻找状态干预项，因此新找到的状态干预项满足：① 具有较大转换效益；② 同已确定的状态干预项关联度小.

5.2.2　群体疾病状态干预

研究适宜于同一类型下病例群体的状态干预方法在生物、医学领域有较大应用价值. 群体疾病状态干预旨在对处于不良状态的病例群体样本挖掘状态干预集进行干预.

文献[9]在个体状态转换项挖掘算法基础上提出群体疾病状态干预挖掘方法，技术要点包括：① 找出每个不良状态病例样本的状态干预集. ② 从所有状态干预项中找出频繁且相关性低的状态干预项组成群体疾病状态干预集.

5.2.3　疾病状态干预评价

考虑临床上无法在短时期内验证状态干预方法的有效性，文献[9]应用分类模型确定被疾病状态干预问题中的干预对象是否处于良好状态，进而评价状态干预的有效性. 若 f 是一个分类器，对象 o 处于不良状态，记为 $f(o)=und$. 设 X 为状态转换集，对象 o 被干预后记为 $X(o)$. 因此，干预 o 的目标是 $X(o)$处于良好状态，记为 $f(X(o))=des$.

目前尚未有任何单个分类模型能完全正确地模拟真实世界，文献[9]采用融合多个分类模型分类结果的方法，来判断被干预对象的状态. 文献[9]在真实数据集上对状态干预算法有效性进行了验证，并通过融合 8 个不同的分类模型评价干预效果，结果表明状态干预方法在大多数实验数据集中都能取得最好的干预效果.

6　未知干预探测研究

6.1　基于二分网的干预发现

在现实世界中，许多网络都呈现出二分结构. 作为复杂网络中一种重要的网络表现形式，二分网由于具有普遍性，被广泛地用于社团结构分析，成为复杂网络研究的重要对象. 为了更好地描述实际应用，可根据实际语义对二分网中节点之间的连接赋予一个权值，构成带权二分网.

文献[10]提出将数据建模成带权二分网，利用二分网社区发现算法找出含有重要信息的社区[11]，再根据发现的社区挖掘干预规则. 文献[10]应用带权二分网研究单因素对出生缺陷发生的影响，并发现干预规则. 这里单因素即出生缺陷监测数据的某一项，如孕妇年龄、文化程度等单个属性.

图 3 示例了孕妇年龄与出生缺陷二分网的示意. 将孕妇年龄进行分组(离散化)，并与各种出生缺陷类型建立二分网，可研究孕妇年龄对出生缺陷发生的影响. 在二分网中，每个年龄组与出生缺陷类型之间连接的权值为该出生缺陷在对应年龄组发生的次数. 通过找出二分网中的年龄社区和出生缺陷社区，可观察出生缺陷发生率在不同年龄社区和出生缺陷社区组合之间的变化，进而发现干预规则.

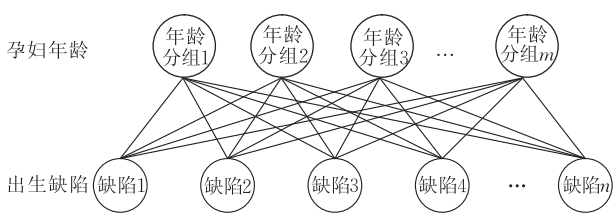

图 3　孕妇年龄与出生缺陷二分网示意

我们团队应用带权二分网分析 1987～1989 年全国出生缺陷监测发现：“中枢神经系统缺陷”，“面、耳和颈部缺陷”，和“肌肉骨骼系统缺陷”，同年龄分

组在 40 岁以上孕妇关联紧密.因此得出对这三类缺陷的干预规则:提倡孕妇在 40 岁之前生产[10].

6.2 流数据环境下干预检测

6.2.1 基于空间数据分布的干预检测

统计分析方法通过对比干预事件发生前后的 ARIMA 回归模型进行干预分析,在现实世界的数据流环境中,一个干预事件可能对数据流中不同数据对象产生不同影响.例如:1997 年暴发的东南亚金融危机,夏威夷总游客数量变化趋势不同于日本游客数量的变化趋势[1].

文献[1]提出一种基于空间数据分布密度,针对高维数据流干预分析的方法.文献[1]指出可以通过观察各个微簇的近似质心与数据点个数来检测干预事件对空间内任意点密度变化的影响.微聚类干预模型对各个微簇的密度吸引特征向量,建立 ARIMA 回归干预模型,进而得到干预变量的回归模型.

文献[1]以 KDD-CUP99 网络入侵数据集中 smurf,back,neptune 3 种不同的网络入侵作为数据流的干预事件进行干预挖掘,得到 3 个不同的干预事件回归模型,通过对比干预事件发生后影响最大的微簇和另一微簇,得到与有关 DoS 网络入侵先验知识一致的实验结论.

6.2.2 单一流数据中干预检测

大量实际应用中的信息都是以数据流的方式获取.若数据流的变化反映事物的发展状态,当没有外界干预存在时,数据流的变化会符合事物发展规律.但当数据流的变化与事物发展方向出现偏差时,预示可能有干预事件的发生.

为了检测数据流中可能的干预事件,文献[4]提出了一种使用配容熵(Entropy Complexity,EC)的方法来自适应检测干预事件.配容熵是热力学的一个概念,最初的用意是描述分子在同一空间不同分割情况下的分布问题.若时间段 T 中窗口 t 的观察数据的项数为 N,将观察数据的取值范围划分为 k 段,每段中数据项数为 $n_i(1\leq i\leq k)$.借鉴配容熵的概念表述数据流中数据项在窗口 t 中的分布情况,计算方法为 $EC(t)=\log(N!/(n_1!*n_2!*\cdots*n_k!))$.这样,时间段 T 内所有观察数据被转换成由一组 EC 值构成的序列 $E(t)$.

对于由数据流转换得到的序列,文献[4]在序列 $E(t)$ 与相邻序列 $E(t-1)$,$E(t+1)$ 之间的离均差(SS)基础上定义序列间显著差异值 $P(t)$,$P(t)=Inf(t)*Inf(t+1)$,其中,$Inf(t)=\max(SS(E(t),E(t+1))/SS(E(t-1),E(t)),SS(E(t-1),E(t))/SS(E(t),E(t+1)))$.

对由 $P(t)$ 构成的观察序列间的显著差异值序列,若 R 为根据实验设定的阈值($R>0$).那么,当显著值超过 R 时,文献[4]认为一个干预事件发生,或干预事件状态发生了改变.

6.2.3 流数据中分段式干预检测

在实际情况下,直接分析整个时序数据流集合,得到的结果往往由于精度过低而失去实际意义.对含有某种周期性规律的数据流,应对数据流的周期性关系分析,才可能得到有意义的结果.为了解决上述问题,文献[7]采用 segment-wise 方式划分原始数据流为可能满足特定分布或规则的子序列,再进行干预分析,检测数据流中是否具有 segment-wise 特性的干预事件.

文献[7]提出数据流稳定性假设:一个数据流系统,若没有外界干预事件发生,系统将保持稳定状态.那么,当系统状态发生偏差时,发生外界干预.

在上述假设基础上,文献[7]设计了挖掘数据流中分段式干预的算法.算法的要点包括:(1)使用 segment-wise 抽取对原始数据集进行处理,并将其划分为彼此满足独立同分布条件的子集,这有助于发现一些微观的数据变化趋势,提高数据挖掘的精度.(2)用指定时间段内泊松参数的函数描述观察数据的特征,将原始数据流转换为特征值序列.(3)将相似的特征值合并,用系统状态描述观察数据的变化趋势.识别初始的系统状态,合并等价系统状态;(4)计算干预事件的影响力,当此影响力超过一定阈值,判定发生了外界干预.

应用分段式干预检测方法,文献[7]在真实金融数据和交通数据上发现了一些有实际用途的事件间隔和未知干预事件,验证了算法的有效性.

6.3 多尺度干预规则挖掘

时间序列反映了事物随时间变化的状态和规律[38].挖掘时间序列间的模式有助于发现序列之间隐藏的联系.因此,干预分析适宜于同时间序列模型结合起来进行研究.大多数时间序列研究将时间序列作为一个不可分割的原子序列.事实上,一个时间序列可能是由若干子序列叠加构成的.在这种情况下,子序列的模式由于序列的叠加而被隐藏,相关系数也无法表达时间序列间各种关系.对此有研究提出时间序列的多尺度分析,即将原时间序列按不同尺度划分为若干子序列,再对子序列进行模式挖掘.以小波为代表的分解技术,被用于时间序列的降维和压缩.

同样,在对时间序列数据进行干预分析时,利用多尺度分析可从不同尺度、不同方向来考虑序列间的干预关系.例如:图 4(a)描述了一位睡眠窒息症患者呼吸频率和心率的时间序列.直接观察图 4(a)很难看出呼吸频率同心率间的干预关系.若提取两

个序列在奇数时间上的子序列，如图 4(b). 容易发现当呼吸频率在第 5 秒提高后，心率紧接着在第 7 秒提高. 可见原始时间序列未明显体现的呼吸频率对心率的干预，清晰体现在奇数时间子序列上.

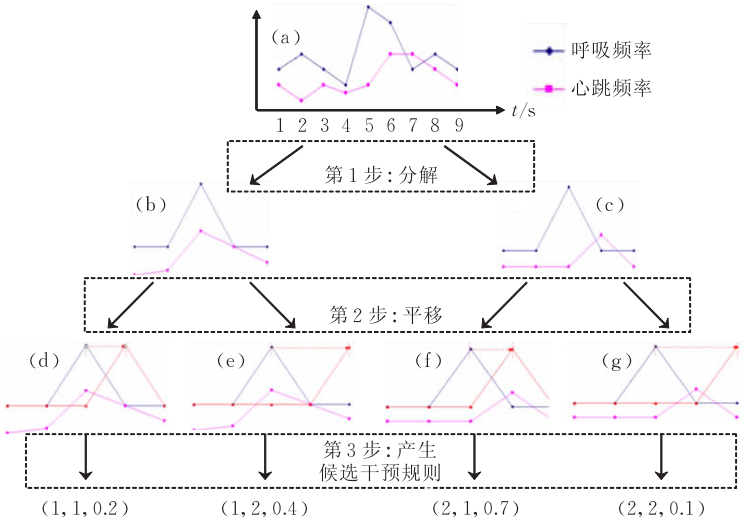

图 4　多尺度干预规则挖掘算法示例

基于上述观察，文献[6]提出了一种采用多尺度分解和相关系数挖掘两个时间序列 X、Y 之间干预规则并评价干预强度的方法. 用三元组 (s,d,i) 表述 X 对 Y 的干预，其中 s 是对原始序列 X 和 Y 的分解尺度(采样频率)，d 是 X 发生变化后 Y 发生相应变化的滞后时间，i 描述 X 对 Y 的干预强度. 下面以挖掘图 4(a)中呼吸频率(X)对心率(Y)的干预规则为例，说明该方法. 图 4 示例了该算法 3 个主要步骤.

1. 分解. 采用小波分解将图 4(a)中的原始时间序列分解成不同采样频率尺度上的两个子序列，分别如图 4(b)、(c)所示. 其中，图 4(b)和图 4(c)分别示例了分解尺度为 1 和 2 时得到的两个子序列.

2. 平移. 由于 X 对 Y 干预的潜在意思是 Y 的变化滞后于 X 的变化，因此将图 4(b)和图 4(c)中 X 的序列延时间轴向右分别平移 1 个和 2 个时间单位，并保持 Y 不变，以表达 Y 的变化相对与于 X 变化的滞后关系. 这样，从图 4(b)的序列得到两种新的 X 和 Y 相对关系，如图 4(d)、(e)所示. 同样，从图 4(c)的序列可得到两种新序列，分别如图 4(f)、(g)所示.

3. 产生候选干预规则集合. 分别计算图 4(d)、(e)、(f)、(g)序列中 X 和 Y 的相关系数 r. 由于这 4 种关系分别表达了 X 变化多长时间后 Y 会发生相似变化，因此把 r 看作 X 发生变化后对 Y 的干预强度.

按照上述步骤分析图 4(a)中的序列可得到 4 条候选干预规则，如图 4 所示. 干预强度的另一种计算方式是采用有向相关系数[12].

文献[6]在真实数据上的进行了实验，结果表明干预规则确实存在于时间序列的不同分解尺度上，并且多尺度挖掘算法比单尺度算法能够发现更多的强干预规则.

6.4　并行事件序列的干预发现

6.3 节介绍了利用多尺度分析法挖掘两个时间序列 X、Y 之间干预规则的方法，在这个基础上，文献[5]将时间序列看作是现实世界中并行事件的发生，考虑在一个时刻可能发生的多个不同类型事件之间的干预. 形式化描述并行事件：设 X 和 Y 是两种不同的事件类型，x_i 和 y_i 分别表示第 i 个周期中 X 类型和 Y 类型的事件个数. 若 y_{i+1} 的分布受 x_i 的影响，说明 X 类型事件在当前周期出现，会影响下一周期 Y 事件的出现，反之不成立，则认为 X 干预了 Y. 亦即，Y 在下一周期的分布不仅依赖于 Y 在当前周期的分布，还依赖于 X 在当前周期的分布.

文献[5]假设事件序列在本质上可用马尔可夫链描述. 记 X 类型和 Y 类型的事件序列分别为 $X(t)$ 和 $Y(t)$，t 为事件发生的时间. 广义马尔可夫性指出，$Y(t)$ 的分布与 $X(t)$ 的分布彼此独立. 但潜在的干预可能破坏广义马尔可夫性. 如果 $P(y_{i+1}\mid y_i)\neq P(y_{i+1}\mid y_i,\ x_i)$，则存在从 $X(t)$ 到 $Y(t)$ 的干预，记作 $X\rightarrow Y$，其中 $P(\cdot)$ 为概率分布函数，称 X 是 Y 的干预源，Y 是 X 的干预目标. 文献[5]证明事件干预 $X\rightarrow Y$ 具有反射性、反对称性和传递性.

在此基础上，将干预效果视为对广义马尔可夫性的偏离，基于 Kullback-Leibler 散度[39]来衡量偏离的程度，称为干预强度[5]，计算如式(1).

$$I(X\rightarrow Y)=\sum P\ (y_{i+1},y_i,x_i)\log\frac{P(y_{i+1}\mid y_i,x_i)}{P(y_{i+1}\mid y_i)} \tag{1}$$

干预强度定量反映了干预源对干预对象的影响. 干预强度的物理意义可以描述为：当实际分布为

$P(y_{i+1}|y_i,x_i)$,而用以编码的假设分布为$P(y_{i+1}|y_i)$时,需要额外使用的比特数.

基于上述讨论文献[5]给出并行事件序列干预发现的基本方法:统计 t 时刻各类型事件的发生次数.对所有类型事件两两组合并计算相互之间的干预强度,保留干预强度大的干预模式,从而得到事件之间的干预模式集合.该方法的时间复杂度为$O(m^2)$, m 为事件类型数量.

7 未来的研究方向

干预规则挖掘是一项新兴的数据分析与决策支持方法.虽然获得了一定的研究成果,但由于应用环境和需求的差异,干预规则挖掘的实现及准确度受此影响,表现出较大的不确定性.

干预规则挖掘在以下方面需进一步探索研究:

(1)干预规则挖掘理论体系研究

目前,有关干预规则挖掘的研究刚起步,理论研究还不深入.目前的研究大多基于特定的应用场景而设计,所提出的方法不具有通用性.由此引出如下问题:在事件序列干预探测中,采用何种评测方法探测干预事件发生最合理?干预强度如何定义?干预规则的描述采用何种标准描述形式等.

同时,建立干预规则挖掘理论体系面临如下挑战:① 干预规则挖掘的需求复杂且不明确.复杂体现在干预规则挖掘过程中可利用的信息有限,面临数据不完整、结构复杂、有效数据少等问题;不明确体现在需求描述主观性强,缺乏量化指标说明需求.② 干预模型建立不统一,针对不同应用建立的干预模型彼此独立,抽取共性十分困难. ③ 现有干预规则挖掘的方法和技术尚不能自成体系.

建立从干预需求到干预模型再到干预规则挖掘技术的理论框架和体系,有望将现有零散方法进行统一,为设计、推广干预规则挖掘技术提供理论支撑和标准参照.

(2)干预动力学模型

干预动力学模型是对干预行为过程的描述与量化.实际应用中,干预措施实施后,干预的效力需要一定时间后才能体现,一次干预作用在持续一定时间后也会逐渐衰减.例如:甲、乙肝疫苗的效力持续时间不相同.可见,不同干预行为的动力学模型有一定差异.

建立干预动力学模型是准确检测干预事件发生、评价干预强度的基础.干预实施后并不一定立即发挥效力,其效力的发挥可能滞后于干预实施一定时间、空间.如何定量和定性描述滞后时间是干预动力学模型建立的要素之一.目前,有关干预强度的研究大多考虑了干预的影响范围,如:受干预作用影响的实例数量等.事实上,干预强度还包括干预作用持续的时间和覆盖的空间等.因此,建立干预动力学模型的要素之二,即是从多方面,定量和定性描述干预的持续效力.

(3)隐性知识对干预规则挖掘的指导

知识管理将高度个性且难于格式化描述的知识称为隐性知识,如主观的理解、直觉和预感.将专家经验用于指导干预规则挖掘是提高干预预测精度,发现高可行性干预规则的重要基础.满足实际需求的干预规则挖掘不能仅参考历史数据,还必须考虑以领域知识为代表的若干相关隐性知识.

将隐性知识转化为易于干预规则挖掘应用的描述是利用隐性知识指导干预规则挖掘的有效途径之一.此外,还涉及隐性知识量化、融合等亟待解决的问题.利用隐性知识评估干预规则还有望提高干预规则的评估精度和干预事件发生检测的准确度.

(4)干预外因分析

在干预规则挖掘过程中,干预外因是产生干预不确定性的因素之一.对未知干预检测,需要判断异常是由干预行为导致的,还是因干预外因引发的;对干预效果预测,需要考虑外界因素对干预效力的影响范围和程度;对干预方法发现,需要考虑干预受外界影响的相应变化,及变化对干预方法的影响.

干预外因是对不确定问题研究的深入.容易想到,现实世界的干预应用都受到各种外界因素影响.干预模型的建立必须考虑干预外因的影响,否则干预模型不具有较好的健壮性和推广性.通过分析干预外因,也能更好地促进干预本质的发现和研究.

(5)实践的检验

目前有关干预规则挖掘的研究大多基于出生缺陷监测数据进行,这既是因为干预规则挖掘问题的提出源于出生缺陷干预,同时也是因为出生缺陷干预是提高出生人口素质的重要手段,意义重大.除此以外,干预规则挖掘还需在其它应用领域进行推广,验证其有效性.这样一方面需要融入具体应用领域的特征,另一方面也为建立完整的干预规则挖掘理论体系,提供了实践检验的条件,有助于高效干预规则挖掘算法的提出和完善,从而推动干预规则挖掘研究和应用的深入开展.

8 结束语

干预规则挖掘是一项源于现实干预应用的新兴数据挖掘任务,旨在利用数据挖掘技术为发现干预

事件，并探索将事物从非期望状态转变为期望状态的方法提供决策支持. 本文概述了干预规则挖掘的基本概念，讲述了同干预规则挖掘相关的其它分析方法，介绍了干预规则挖掘的基本流程及任务分类，并围绕干预规则挖掘的分类任务，叙述了干预规则挖掘的主要研究问题及相关进展，最后总结全文，并指出干预规则挖掘未来的研究方向.

致　谢　本文在完成过程中得到了多位曾经在四川大学计算机学院数据库与知识工程研究所学习过的毕业生的支持，在此表示感谢！

参　考　文　献

[1] Tang Chang-Jie, Zhang Yue, Tang Liang, Li Chuan, Chen Yu. Survey on mining kinetic intervention rule from sub-complex systems. Journal of Computer Applications, 2008, 28(11): 2732-2736; 2748(in Chinese)
(唐常杰，张悦，唐良，李川，陈瑜. 亚复杂系统中动力学干预规则挖掘技术研究进展. 计算机应用，2008，28(11)：2732-2736；2748)

[2] Tang L, Tang C J, Duan L, Li C, Jiang Y X, Zeng C Q, Zhu J. MovStream: An efficient algorithm for monitoring clusters evolving in data streams//Proceedings of the 2008 IEEE International Conference on Granular Computing. Hangzhou, 2008: 582-587

[3] Zhang Yue, Tang Chang-Jie, Li Chuan, Zhu Jun, Zeng Chun-Qiu, Tang Liang, Liu Xian-Bin. Mining Naïve intervention rules in birth defect data. Journal of Frontiers of Computer Science and Technology, 2009, 3(2): 188-197(in Chinese)
(张悦，唐常杰，李川，朱军，曾春秋，唐良，刘显宾. 出生缺陷监测数据中的朴素干预规则挖掘. 计算机科学与探索，2009，3(2)：188-197)

[4] Wang Y, Tang C J, Li C et al. Intervention events detection and prediction in data streams//Proceedings of WAIM 2009. Suzhou. LNCS 5446. 2009: 519-525

[5] Yang N, Tang C J, Wang Y. Mining interventions from parallel event sequences//Proceedings of the WAIM 2009. Suzhou. LNCS 5446. 2009: 297-307

[6] Zheng J L, Tang C J, Qiao S J, Yang N, Wang Y, Chen Y, Zhu J. MMIR: Mining multi-scale intervention rules in sub-complex system//Proceedings of the APWeb 2010. Busan, 2010: 369-371

[7] Wang Y, Zuo J, Yang N, Duan L, Li H J, Zhu J. An efficient approach for mining segment-wise intervention rules in time-series streams//Proceedings of the WAIM 2011. Jiuzhaigou. LNCS 6184. 2011: 194-205, 2010

[8] Tang Chang-Jie, Duan Lei, Wang Yue, Yang Ning, Zhu Jun, Dai Li. Task classification of intervention rules mining and advances of three technologies. Journal of Computer Applications, 2010, 30(1): 10-14(in Chinese)
(唐常杰，段磊，王悦，杨宁，朱军，代礼. 干预规则挖掘的任务分类和三项技术进展. 计算机应用，2010，30(1)：10-14)

[9] Dong G, Duan L, Tang C. Mining disease state converters for medical intervention of diseases. Journal of Bioinformatics and Computational Biology, 2010, 8(1): 77-97

[10] Xu Kai-Kuo. The research for discovering communities from bipartite network and its application on Interventional rule mining[Ph. D. dissertation]. Sichuan University, Chengdu, 2010(in Chinese)
(徐开阔. 二分网社区发现技术及在干预规则挖掘中的应用[博士学位论文]. 四川大学，成都，2010)

[11] Xu K K, Tang C J, Li C, Jiang Y X, Tang R. An MDL approach to efficiently discover communities in bipartite network//Proceedings of DASFAA 2010. Tsukuba, 2010: 595-611

[12] Zheng Jiao-Ling. Mining evolutionary characteristics from sub-complex system via dynamic intervention [Ph. D. dissertation]. Sichuan University, Chengdu, 2010(in Chinese)
(郑皎凌. 动态干预条件下的亚复杂系统演化特征挖掘[博士学位论文]. 四川大学，成都，2010)

[13] Tang Chang-Jie, Duan Lei, Zheng Jiao-Ling, Yang Ning, Wang Yue, Zhu Jun. Mining causality, segment-wise intervention and contrast inequality based on intervention rules. Journal of Computer Applications, 2011, 31(4): 869-873(in Chinese)
(唐常杰，段磊，郑皎凌，杨宁，王悦，朱军. 基于干预规则挖掘因果关系与分段干预事件及对照不等式. 计算机应用，2011，31(4)：869-873)

[14] Wang Yue, Tang Chang-Jie, Yang Ning, Zhang Yue, Li Hong-Jun, Zheng Jiao-Ling, Zhu Jun. Mining optimized probabilistic intervention strategy over uncertain data set. Journal of Software, 2011, 22(2): 285-297(in Chinese)
(王悦，唐常杰，杨宁，张悦，李红军，郑皎凌，朱军. 在不确定数据集上挖掘优化的概率干预策略. 软件学报，2011，22(2)：285-297)

[15] Duan Lei, Zuo Jie, Li Chuan, Chen Yu, Tang Chang-Jie, Zhu Jun, Dai Li, Mou Xin. HealthyBaby: A mining system over birth defects data of China. Journal of Computer Research and Development, 2010, 47(z1): 520-534(in Chinese)
(段磊，左劼，李川，陈瑜，唐常杰，朱军，代礼，牟昕. 中国出生缺陷数据挖掘系统 HealthyBaby. 计算机研究与发展，2010，47(z1)：520-534)

[16] Box G E P, Tiao G C. Intervention analysis with applications to economic and environmental problems. Journal of the American Statistical Association, 1975, 70(349): 70-79

[17] Xiong Yan, Yu Shi. Application of intervention analysis by ARMAX. Statistics and Decision, 2003, (11): 17-26(in Chinese)
(熊焰，余石. 干预分析的 ARMAX 模型及应用. 统计与决策，2003，(11)：17-26)

[18] Zhu Y Y, Shasha D. Efficient elastic burst detection in data streams//Proceedings of the SIGKDD 2003. Washington, 2003: 336-345

[19] Ihler A, Hutchins J, Smyth P. Adaptive event detection with time — Varying poisson process//Proceedings of the SIGKDD 2006. Philadelphia, 2006: 207-216

[20] Lakshmanan L V S, Russakovsky A, Sashikanth V. What-if OLAP queries with changing dimensions//Proceedings of the ICDE 2008. Cancun, 2008: 1334-1336

[21] Wang Shan, Xiao Yan-Qin, Zhang Yan-Song, Chen Hong. Research on OLAP system supporting What-if analysis. Chinese Journal of Computers, 2008, 31(9): 1573-1586(in Chinese)
(王珊, 肖艳芹, 张延松, 陈红. 支持 What-if 分析的 OLAP 系统研究. 计算机学报, 2008, 31(9): 1573-1586)

[22] Zhang Yan-Song, Xiao Yan-Qin, Wang San, Chen Hong. What-if query processing policy of main-memory OLAP system. Journal of Software, 2010, 21(10): 2494-2512(in Chinese)
(张延松, 肖艳芹, 王珊, 陈红. 主存 OLAP 系统中 what-if 查询处理策略. 软件学报, 2010, 21(10): 2494-2512)

[23] Xiao Y Q, Zhang Y S, Wang S, Chen H. Efficient incremental computation of CUBE in multiple versions what-if analysis//Proceedings of the WAIM 2009. Suzhou. LNCS 5446. 2009: 235-247

[24] Zhang Y S, Zhang Y, Xiao Y Q, Wang S, Chen H. The tradeoff of delta table merging and re-writing algorithms in what-if analysis application//Proceedings of the WAIM 2009. Suzhou. LNCS 5446. 2009: 260-272

[25] Golfarelli M, Rizzi S. Designing what-if analysis: Towards a methodology//Proceedings of the International Workshop on Data Warehousing and OLAP. Arlington, 2006: 51-58

[26] Golfarelli M, Rizzi S. What-if simulation modeling in business intelligence. International Journal of Data Warehousing and Mining, 2009, 5(4): 24-43

[27] Wang K, Jiang Y, Tuzhilin A. Mining actionable patterns by role models//Proceedings of the ICDE 2006. Atlanta, 2006

[28] Yang Q, Cheng H. Mining plans for customer-class transformation//Proceedings of the ICDM 2003. Melbourne, 2003: 403-410

[29] Yang Q, Yin J, Ling C X, Pan R. Extracting actionable knowledge from decision trees. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(1): 43-56

[30] Li J, Wong L. Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns. Bioinformatics, 2002, 18(5): 725-734

[31] Dong G, Li J. Efficient mining of emerging patterns: Discovering trends and differences//Proceedings of the KDD 1999. San Diego, 1993: 43-52

[32] Liu B, Hsu W, Han H S, Xia Y. Mining changes for real-life applications//Proceedings of the DaWaK 2000. London, 2000: 337-346

[33] Wang K, Zhou S, Fu A, Yu J X. Mining changes of classification by correspondence tracing//Proceedings of SDM 2003. San Francisco, 2003

[34] Jiang B, Pei J, Lin X, Cheung D W L, Han J. Mining preferences from superior and inferior examples//Proceedings of KDD 2008. Las Vegas, 2008: 390-398

[35] Ferreria C. Gene Expression Programming Mathematical Modeling by an Artificial Intelligence. Berlin: Springer-Verlag, 2006

[36] http://www.ornl.gov/sci/techresources/Human_Genome/medicine/genetherapy.shtml

[37] http://www.cancer.gov/cancertopics/factsheet/Therapy/gene

[38] Han J, Cheng H, Xin D, Yan X. Frequent pattern mining: Current status and future directions. Data Mining and Knowledge Discovery, 2007, 15(1): 55-86

[39] Kullback S, Leibler R A. On information and sufficiency. The Annals of Mathematical Statistics, 1951, 22(1): 79-86

DUAN Lei, born in 1981, Ph. D., lecturer. His research interests include data mining, evolutionary computation.

TANG Chang-Jie, born in 1946, professor, Ph. D. supervisor. His research interests include database, data mining and knowledge engineering.

YANG Ning, born in 1974, Ph. D., lecturer. His research interests include data mining and machine learning.

ZUO Jie, born in 1977, Ph. D., lecturer. His research interests include database and data mining.

WANG Yue, born in 1981, Ph. D.. His research interests include database and data mining.

ZHENG Jiao-Ling, born in 1981, Ph. D., lecturer. Her research interests include database and knowledge engineering.

XU Kai-Kuo, born in 1983, Ph. D., lecturer. His research interests include evolutionary computation and data mining.

Background

Intervention rule mining is derived from the practice in the real intervention application, such as birth defects intervention. The targets of intervention rule mining include: intervention event detection, the best intervention time and intensity discovery, and decision support for converting objects from undesirable state to desirable one. The intervention rule mining has a wide range of applications in the real world. In this paper, the authors survey the research issues, difficulties and achievements of the intervention rule mining, also discusses the future research direction of intervention rule mining.

The work is supported by the National Natural Science Foundation of China (grant Nos. 60773169, 61103042, 61173099), the 11th Five Years Key Programs for Science & Technology Development of China (grant No. 2006BAI05A01), the Doctoral Foundation of Ministry of Education of China (grant No. 20100181120029), and the Young Faculty Foundation of Sichuan University (grant No. 2009SCU11030). During the past four years, the database and knowledge engineering institute of School of Computer Science, Sichuan University, had performed quite a few researches on intervention rule mining, as well as implemented a mining system over birth defects data of China.