文章编号:1001 - 9081(2008)11 - 2732 - 05

亚复杂系统中动力学干预规则挖掘技术研究进展

唐常杰,张 悦,唐 良,李 川,陈 瑜
(四川大学 计算机学院,成都 610064)
(tangchangjie@cs.scu.edu.cn; chjtang@vip.sina.com)

摘 要:亚复杂系统干预规则挖掘是数据挖掘领域的新内容。综述了亚复杂系统干预规则研究背景和典型问题,通过实例,描述了干预规则挖掘领域一些基本概念和术语,如干预相关度、传递相关度、干预分型和干预代数等;介绍了在亚复杂系统干预规则挖掘的初步探索和成果,包括关于朴素干预规则和数值型干预规则挖掘算法,以及基于密度的数据流干预分析模型及相关结果。

关键词:数据挖掘;亚复杂系统;干预规则;出生缺陷
中图分类号: TP311.13 **文献标志码**:A

Survey on mining kinetic intervention rule from sub-complex systems

TANG Chang-jie, ZHANG Yue, TANG Liang, LI Chuan, CHEN Yu
(*College of Computer Science, Sichuan University, Chengdu Sichuan* 610064, *China*)

Abstract: Intervention rule Mining from Sub-Complex system (IMSC) is a new hot spot in data mining area. The background and typical problems in IMSC were surveyed. The related concepts and terminologies in IMSC were described, such as intervention correlation, intervention type and intervention algebra, the authors' research results were briefly introduced, including naïve rules mining, numerical intervention rules mining, and data stream intervention analysis based on density as well.

Key words: data mining; Sub-Complex System (SCS); intervention rule; birth defect

0 引言

现实世界中的一大类困难问题属于复杂系统问题,复杂系统有三个宏观特征: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[1-4]。3)简构复行,即由简单对象构成却有复杂行为表现。复杂系统元素微观上相互作用、自我调整,而系统在宏观上表现为生物、经济、社会等复杂系统中的平衡、调整和剧变。复杂系统有四大特色:1)内部自适应性,系统行为有良好规律,可感知环境、调整行为,可产生从未有过的新规则,使系统更有序。2)局部性,系统对象不能感知和控制非局部其他对象的状态行为。3)竞争性,系统整体行为是个体竞争、协作的综合;4)突变性、非线性和不确定性。

复杂系统研究有两大学派:1)圣塔菲(Santa Fe)学派,以计算机模拟对象在虚拟环境下相互作用,自下而上表现系统复杂性行为。2)控制论学派,自上而下地通过有限理性和不确定信息做出合理决策。由于复杂系统自身的这些特点,目前的技术很难在真正意义上有效地解决复杂系统领域的问题。

亚复杂系统对复杂系统进行特征提取、忽略次要因素、降维等处理,得到一个较简单、且目前科学能力有可能做出工程性解决方案的系统,简称亚复杂系统(Sub-Complex System, SCS)。

亚复杂系统干预动力学规则描述人工干预下亚复杂系统的动力学行为,揭示干预的普适规律,并提出可工程性化的解决方案。

由此延伸出一系列复杂应用,如人工增雨、人工防雹、出生缺陷干预、糖尿病干预、地壳断裂带能量的局部受控释放(如深井人工地震等)、堰塞湖能量的受控释放等。这些应用涉及了自然科学、工程科学和社会科学等广泛领域的问题,是亚复杂系统的典型研究背景。为了形象、直观地给出其内涵和外延,本文通过亚复杂系统干预的一些典型实例来对其进行描述。

1 SCS 干预问题实例

实例 1 出生缺陷干预 我国 20 年出生缺陷监测数据表明[5-8]:1)每年因神经管缺陷、唐氏综合症和先天性心脏病引起的经济损失达到 200 亿元。2)出生缺陷在经济贫穷落后地区发生率高。3)出生缺陷存在病种、地区、人群间的差异。全国出生缺陷发生状况已积累近数十年连续监测、干预历史数据,在这些数据中进行干预动力学规律研究、挖掘不同地区及不同干预方式的效果、分析影响干预效果的医学和社会学因素等行为将为决策提供依据。

实例 2 糖尿病干预 研究者从复杂系统的角度对糖尿病控制进行了研究[9-12],糖尿病控制是亚复杂系统干预特例。糖尿病亚健康状态的发展有如下特点:1)样本量较大,通常涉及万人几十年数据。2)维数较多(通常多于 100 维),如身高、体重、年龄、血糖值、血压值等,含主观属性,如情绪、适应等。3)发病规律具地域性、群体性特点。4)任务复杂。糖尿病控制包括亚健康状态发生相关因素作用程度、常见发病危险因素的临界值、发病危险因素分层结构、重要发病危险因素的提取、重

收稿日期:2008 - 07 - 08。 **基金项目**:国家自然科学基金资助项目(60773169);"十一五"国家科技支撑计划项目(2006BAI05A01)。
作者简介:唐常杰(1946 -),男,重庆人,教授,博士生导师,主要研究方向:数据库与知识工程、数据挖掘; 张悦(1983 -),女,四川自贡人,硕士研究生,主要研究方向:数据挖掘; 唐良(1983 -),男,重庆人,硕士研究生,主要研究方向:数据挖掘; 李川(1977 -),男,河南郑州人,讲师,博士,主要研究方向:数据挖掘; 陈瑜(1974 -),男,陕西汉中人,讲师,博士,主要研究方向:数据挖掘。

要发病危险因素的作用程度分析,以及发病状态预测及宏观规律等诸多因素。这些因素呈现出多源、多型、多精度、多时态、多维非线性交错等特征,使系统干预异常复杂。

实例 3　市场调控　例如,房价市场的政策性干预问题。影响房价的因素很多,如竣工造价、当地人均可支配收入、地方税收、行政费用、自然灾害、国际经济形势等。政策性干预是保障房地产健康发展的必要方式。决策部门向有关专家提出下列问题:采用何种手段、多大力度等措施方能使房屋市场价格康复到何种程度?

实例 4　扩招与就业　大学扩招有积极的社会意义,能推迟就业压力峰值,但同时产生负面影响,脉冲式的教师负担和生源质量下滑等,埋伏下几年后的另一次就业压力峰值。采用何种政策或技术手段,多大的力度,能得到多大的控制效果,在几年后有多大效益? 能否从数据中挖掘(而不是主观臆造)出一个精确到二阶微分(或差分)的,能描述扩招——延缓——下一次峰值的动力学规律? 这是关于亚复杂系统干预规则问题。

此外,亚复杂系统干预领域问题在现实生活中还有很多实例,如 10 万人以下的小型人群流动的干预、微型生态系统的干预等。

亚复杂系统干预动力学研究涉及分类、预测,关联和多维数据分析、公式发现等基本数据挖掘技术,但绝不是这些技术的简单叠加。为有效解决此类问题,作者提出新的数据挖掘任务——"系统干预规则"挖掘,简称"干预规则"挖掘。数据挖掘中,它是与关联、聚类、预测等传统任务平行,甚至更高级别的任务。亚复杂系统的干预规则从复杂问题中剥离混沌性质,萃取出亚复杂系统,挖掘亚复杂系统在人工干预的动力学行为、揭示干预普适规律。

2　SCS 干预模型

本章以出生缺陷干预为例,进一步探讨 SCS 干预模型。

设 $A_t=(a_1, a_2, \cdots, a_n)$ 记录时间 $t(t>0)$ 时观察的属性值,其中属性 $a_i(1\leqslant i\leqslant n)$ 与特定地域出生缺陷相关,包括生育年龄、地区、文化水平、营养状况、叶酸补充、婚检率等因素。设 $B_j(0<j\leqslant m)$ 是最常见的 m 种出生缺陷的发生率。G_t 是在 t 时刻采用的政策和技术干预措施的量化指标。在干预实施周期内观察数据 C_t, = {干预措施 G_t:实施代价 D_t}。20 年来中国人口出生缺陷的观测数据在对象关系模式(A_i, B_j, C_t) 上组成含有数百万条记录的稀疏矩阵。

在进行出生缺陷干预决策时,应该在何种条件下、采用多大的干预投资 G(亿元) 等策略方可得到多大干预效益(如缺陷率低于某阈值等)?此问题可泛化为规则 1。

规则 1　亚复杂系统 S 中,对于给定一组阈值$\{g_k,\varepsilon,\lambda\}$,$\{G_k>g_k\mid,0<k<120\}\rightarrow\{B_j<\lambda\&\&(\delta y_j/\delta t)<\varepsilon(0<j\leqslant 100\}$ *supp*, *Conf*。

上述规则成立的含义是:当干预力度 $G_k>g_k$ 时能保证出生缺陷发生率 $y_j<\lambda$,并且出生缺陷发生率的偏导数(年增加率)$(\delta B_k/\delta t)<\varepsilon$。同时,在测试数据集上满足支持度大于阈值 *supp*,置信度大于阈值 *Conf*。其中的一阶或高阶偏导数(或差分) 反映了系统动力学行为。

规则 Ic 称为亚复杂系统的 I 型干预规则。本文回避复杂公式和定义,将用实例来说明干预规则的意义。

例 1　表 1 给出 A 地区 1995 到 2000 年的"年龄—叶酸—缺陷"数据片段。其中"叶酸不足且新生儿缺陷"的记录占 2/6,"叶酸充足且新生儿缺陷"的记录占 1/6。

表 1　孕母年龄—补充叶酸—新生儿缺陷关系

编号	年龄	叶酸量	有无缺陷
1	38	足	有
2	40	不足	有
3	36	不足	有
4	24	足	无
5	30	足	无
6	28	不足	无

干预理论认为,当叶酸状态由"不足"干预为"充足"时,原先叶酸不足的人(包括缺陷和健康)中,有 1/3 的人会转为健康,这个人数占原"叶酸不足且新生儿缺陷"人数的 1/2。把这个结果表达为干预规则式为: Location ("A") ∧ Time ("1995 - 2000") | 叶酸(不足→足) ⇒有无缺陷(有→无) 支持度 = 2/6,变化度 = 1/3,置信度 = 1/2。

本文引入一系列概念,如干预属性、朴素干预属性、数值型和混合型干预规则,影响因子 ψ(衡量被干预属性的响应),支持度阈值 *supp*,置信度阈值 *Conf*。其中 ψ 反映系统动力学行为。

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、离散型、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。综上所述,亚复杂系统干预研究有下列两个典型特征:

1)实践性。亚复杂系统干预是从出生缺陷干预、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。在进行 SCS 干预研究时,应简化问题、萃取出亚复杂系统,从而有效降低干预成本、提高干预效益。

2)基础性。亚复杂系统干预规则挖掘由一系列基本挖掘技术组成,如分类、聚类、预测、关联和多维数据分析等。干预规则挖掘作为新的挖掘任务,必将吸引研究者的热烈参与,成为一个新的创新源头和学术园地。

3　SCS 干预研究的内容

本文以出生缺陷干预数据为具体对象,萃取出亚复杂系统,挖掘隐藏于其中的干预规则,并对 SCS 干预行为进行动力学分析。包括五方面研究内容:

1)干预相关度和传递相关度。现实世界的 SCS 包含多个不同的、相关的属性,增加了分析某属性干预效果的难度。例如,出生缺陷干预的数据集合中,文化程度高常导致出生缺陷低,但文化程度高又常导致生育年龄高。生育年龄高又导致出生缺陷高等。实践表明,相关度大的干预适宜进行并发干预研究。相关度小的干预可分离到不同的亚复杂系统中,降低问题难度。

2)SCS 的萃取。从复杂系统中分离出 SCS,可能的途径包括:①降维,可采用信息熵、JINI 指数等方法或提出新方法来分离出重要维度。②剥离相关度小的干预。

以出生缺陷分析为例,要从 200 维、50 个并发干预的复杂系统$\{X_1, X_2, \cdots, X_{200}, G_1, G_2\cdots, G_{50}\}$ 中萃取出不超过 100 维,并且不超过 20 个并发干预手段的亚复杂系统$\{x_1, x_2, \cdots, x_{100}, g_1, g_2\cdots, g_{20}\}$。同时,保证亚复杂系统的动力学核$\{(\delta g_k/\delta t), \delta g_k/\delta x_j\}$ 的精度能适应领域需求。

3)干预分型和干预代数。干预规则分为多种类型。描述动力学性质的常常涉及微分(或差分)、描述串联干预的为管道式规则 G_1 干预 G_2、G_2 干预 G_3 以及包括次序敏感型和次序

不敏感型等的并连的干预行为。本文提出干预代数概念,它包括一个系统、一组干预、一组规则等。

4)干预规则的挖掘。干预规则挖掘难度高于传统关联规则挖掘,体现在:①0阶训练数据、一阶数据或高阶差分数据挖掘。②临界阈值的挖掘。临界阈值在糖尿病干预或出生缺陷干预中提示最佳干预时机,有最小的损失,最大的收益。③基于基因表达式编程干预规则的挖掘。包括亚复杂系统的特征如何用基因表达式编程来描述、如何用适应度函数指挥干预规则的进化?它的动力学规则需要什么条件?如何优化?收敛性性质如何等?

5)干预的解释和评价。干预规则在微分或差分形式下描述系统微观状态即可回答关于系统行为的"为什么",进一步帮助理解干预机制、改善干预技术或提供干预咨询依据。为此,需对产生的多个候选干预规则,建立合理的评价体系。

4 我们的工作

SCS人工干预是一个挑战性问题,限于篇幅,这里仅介绍已进行的朴素干预规则和数值型干预规则研究。

4.1 朴素干预规则挖掘

旨在找出在一定程度上反映因果关系的干预规则,反映出干预前后的变化。关联规则能够发现大量潜在的关联信息和知识,以反映事务间的关联关系的方式为人们提供决策帮助,可以用于上述问题中的相关挖掘。为了使关联规则以反映最佳干预方向的方式提供决策建议,本文在规则中对前件属性与后件属性的变化量关系进行描述,建立了朴素干预规则模型,在其中提出并定义了变化度Delta度量,并用其来反映干预效果。

定义1 朴素干预规则:设数据集 D 的属性集 $A = \{A_1, A_2, \cdots, A_w\}$,对所有 $A_i \in A$,设 $\{|A_i|\}$ 表示 A_i 的取值或状态集合。设用户选定的干预行为的自变属性集 $AA = \{AA_1, AA_2, \cdots, AA_m\}$,用户选定的因变属性集 $DA = \{DA_1, DA_2, \cdots, DA_n\}$,$AA \cap DA = \varnothing$ 且 $AA \cup DA \subseteq A$。满足下列条件的表达式成为 r 朴素干预规则:

1)存在集合 $AA' \subseteq AA$,对所有 $AA_p \in AA$,存在 $aa_i \in \{|AA_p|\}$,项 $AA_p(aa_i)$ 表示为 I_p';存在 $DA_q \in DA$,且 $da \in \{|DA_q|\}$,项 $DA_p(da)$ 表示为 I_q';使得关联规则 $r': I_1' \wedge I_2' \wedge \cdots \wedge I_{|AA'|}' \Rightarrow I_q'$ 为强关联规则,支持度为 $Sup(r')$,置信度为 $Conf(r')$。

2)存在 $aa_j \in \{|AA_p|\}$,且 $aa_i \neq aa_j$,项 $AA_p(aa_j)$ 表示为 I_p'',关联规则 $r'': I_1'' \wedge I_2'' \wedge \cdots \wedge I_{|AA'|}'' \Rightarrow I_q'$,支持度为 $Sup(r'')$,置信度为 $Conf(r'')$。

3)项 $AA_p(aa_i \rightarrow aa_j)$ 表示为 I_p,项 $DA_p(da \rightarrow !da)$ 表示为 I_q,则 $r: I_1 \wedge I_2 \wedge \cdots \wedge I_{|AA'|} \Rightarrow I_q$。

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

$$\begin{cases} Sup = Sup(r') \\ Delta(r) = Conf(r') - Conf(r''),\ Delta(r) > 0 \\ Conf(r) = Delta(r)/Conf(r') \end{cases} \quad (1)$$

直观地解释,变化量 $Delta(r)$ 体现了前件的变化引起后件变化的方向和幅度,当 $Delta(r)$ 大于0时规则前件和后件是正相关的。此外:1)规则 r 的支持度 $Sup(r) = Sup(r')$,是干预所影响的范围大小的体现。2)置信度 $Conf(r) = Delta(r)/Conf(r')$,表明了变化量占原始状态情况的百分比,是变化准确率的体现。

朴素干预规则挖掘算法由三个部分组成:

1)按文献[13-14],将出生缺陷关系型数据库转换为事务数据库;

2)挖掘频繁项集,采用Apriori算法或FP-growth算法[15-16],规则 r' 后件由人工指定;

3)扫描频繁项集,计算同类属性事务关联规则 r,挖掘出 $Delta$ 值最大的项组成朴素干预规则 r。

我们在朴素干预规则挖掘方面做了一系列实验。实验平台为{AMD Athlon 1.61 GHz,512 MB内存,Windows XP,BC++ Builder 6.0}。实验数据为1986年至1991年的全国围产儿数据(由全国出生缺陷检控中心提供),共200万条记录;进行了朴素干预规则分析处理,实验结果中得到三条指导意义的规则:

规则2 时间(1986-1987)|近亲结婚(是→否)⇒缺陷儿(是→否) $Sup = 0.0240$, $Delta = 0.3388$, $Conf = 0.9906$。

规则3 时间(1986-1991)|城乡(乡村→城镇)∧缺陷儿(是→否)⇒围产儿死亡(是→否) $Sup = 0.0118$, $Delta = 0.8284$, $Conf = 0.9832$。

规则4 时间(1986-1991)|,时间(1986-1987)|先天患病(有→无)⇒缺陷儿(是→否) $Sup = 0.0005$, $Delta = 0.4622$, $Conf = 0.9307$。

规则2表明,"近亲结婚"为"是"时缺陷率较高。推荐把"近亲结婚"干预成"无",可能会有33.88%的人不会出现缺陷。规则3表明,1)农村的缺陷儿死亡率较城镇的高。2)若要降低死亡率,推荐干预为城镇的非缺陷儿。此外,由于城乡地域在现实生活中是无法干预的,则该规则只能起到发现相关性的作用。规则4表明,患有先天疾病的父母后代为缺陷儿的比例较大,推荐减少患有先天疾病的父母。

表2给出了在频繁项集数量基本相同时运行算法第三部分所得到的数据量与运行时间的对应情况,可见算法第三部分的伸缩性较好。

表2 算法第三部分的伸缩性

数据量	运行时间/s
0	0
18624	1
500000	15
1000000	36
1500000	55
1800000	70
1980000	80

可见朴素干预规则能对分析诱发新生儿缺陷的因素提供一些难能可贵的、有指导意义的素材,从而为决策者制定出生缺陷干预措施方案提供一定的理论基础和有效的决策依据。

4.2 数值型干预规则挖掘

上述三朴素干预规则方法是基于范畴型(非数值性)数据。为了满足实际应用,作者提出了基于数值型属性的干预规则挖掘算法。

数值型干预规则挖掘算法由三个部分组成:

1)判断自变属性和因变属性是否相关;

2)对相关属性集,得到模型的关系函数 f;

3)分析 f 的单调性,分区间计算支持度、变化度和置信度。

其中:1)把数据集分为两组,分别进行参数回归,得到两个函数,比较它们的泰勒公式系数,如果差异大于一个给定的

误差阈值,则认为考查的数值属性是不相关的;反之,则认为这两个函数“大概相同”,考查的数值属性是相关的。2)中,选择两个函数中较好者作为数值型干预规则的函数;3)中,为了指导干预方向,需要划分函数的单调区间。使用一阶导数判断法即可。各单调区间的单调性用 ASC 或 DESC 表示。

干预规则挖掘所面对的训练数据的特点是,数据集中自变量相同的值,因变量的值也可以不同,而且可以有多个相同的数据。由于数据的特殊性,使用 GEP 来做非线性主成分分析。文献[17]提出的“弱适应模型”可以用来解决非线性主成分的挖掘,该模型使用“带内集”来避开噪声数据的影响。

数值型干预规则中,作者在传统的支持度、变化度和置信度等度量参数上做了一定的改进,区间的支持度反映了该区间的重要性;变化度为干预关系函数的一阶导数,反映了函数变化的趋势;置信度是成为了主成分的数据个数占所有个数的百分比。例如,当数值型干预规则主成分函数的单调性有两种情况的时候,该干预规则可以表示为:$A \in (a1,a2):(a2,a3) \Rightarrow B = f(a)$ $mono1:mono2$ $supp1:supp2$, $delta$, $conf$。

作者做了数值型干预规则挖掘实验。实验环境和数据与朴素干预规则实验相同。GEP 参数设置与文献[18]的表 5.4 相同,带宽为 2.000。在围产儿月报表和季报表上进行挖掘,数据量为25 000条,以下是实验结果中的其中两条规则:

规则 5　时间($a \in [86,91]$)⇒围产儿死亡率($-0.872a + 102.330$)

DESC $supp = 1, delta = -0.872, conf = 1$。

规则 6　围产儿死亡率(单位:‰)($a \in [22.2,26.7]$)⇒死亡中缺陷率(单位:%)($-1.456a^2 + 65.874a - 697.730$)

DESC $supp = 1, delta = -1.456a + 65.874, conf = 1$。

规则 5 表明,围产儿死亡率随时间而减少,由于时间不可被干预,本规则只起到了发现相关性的作用。规则 6 表明,当死亡率下降时,死亡中的缺陷率增加,则说明如果降低缺陷率,可能降低死亡率,这与朴素干预规则在围产儿基本信息上得到的实验结果一致。

图 1 给出了在 5 个数值属性上成功训练出模型所涉及到的数据量和运行时间对应情况,可见用 GEP 进行模型训练具有稳定的伸缩性。

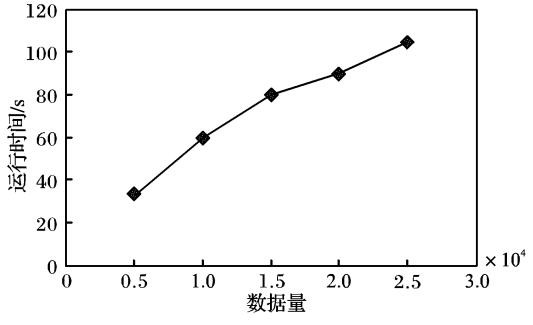

图 1　算法的伸缩性

可见数值型干预规则对分析新生儿统计信息有很大作用,它能挖掘出具有相关性的数值属性,决策者可以参考单调性做出决策。

4.3　基于密度的数据流干预分析

传统时间序列数据干预分析模型通过挖掘干预事件发生前后两个 ARIMA 回归模型做分析[13],适用于统计量数据的分析,在现实世界数据流环境下,一个干预事件可能对数据流中不同数据产生不同的影响效果。以 1997 年爆发的东南亚金融危机为例,对夏威夷旅游产业进行干预分析。数据表明,东南亚金融危机对日本、新加坡等东亚游客的影响远大于北美、欧洲等地区游客的影响。表 3 显示来自夏威夷的官方统计数据,1997 ~ 2000 年,夏威夷总游客数量变化趋势与日本游客数量的变化趋势是不相同的。

表 3　1995 ~ 2000 年夏威夷游客统计

年	总的游客数量	日本游客数量
1995	6 546 759	2 048 411
1996	6 723 141	2 146 883
1997	6 761 135	2 216 890
1998	6 595 790	2 004 354
1999	6 741 037	1 825 587
2000	6 948 595	1 817 644

我们提出一种基于空间数据分布密度,针对高维数据流上干预分析的方法。空间内的任一位置 X 的密度通过高斯 Kernel 密度估计方法计算而得[20],如下式:

$$\hat{f}(X) = \frac{1}{nh\sqrt{2\pi}} \sum_{i=1}^{n} e^{-(X-X_i)^2/(2h^2)} \tag{2}$$

其中$\hat{f}(X)$ 为空间中 X 处的估计密度,n 为数据点个数,X_i 为第 i 个数据值,$i = 1,2,\cdots,n$, h 为平滑参数。

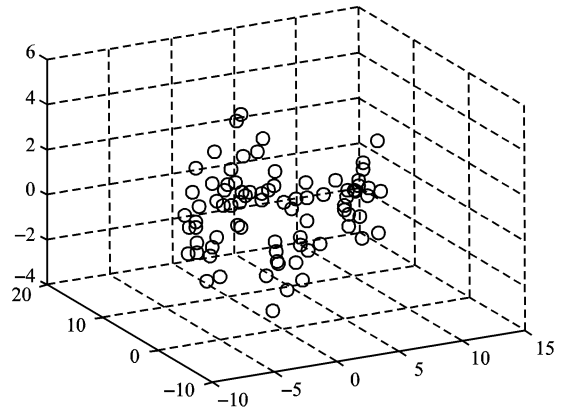

图 2　3 维空间内的微聚类结果

我们在针对高维数据流上的干预分析方法主要包括实时的微聚类(Micro-Clustering)与非实时的微聚类干预模型建立两个过程。图 2 是 3 维空间内的微聚类的例子。为清楚表达,引入下列的术语和记号:

给定一个微聚类簇 c_i, $s(c_i)$ 为 c_i 内的数据点集,$n(c_i)$ 为簇内的数据点个数,$c(c_i)$ 为 c_i 的质心,$c(c_i) = \sum_{X_j \in s(c_i)} X_j / n(c_i)$,$r(c_i)$ 为 c_i 的半径,$r(c_i) = \left(\frac{1}{n(c_i)} \sum_{X_j \in s(c_i)} | c(c_i) - X_j |^2\right)^{1/2}$,$M$ 为所有微簇集合。ε 为用户给定阈值,且为允许微簇半径的最大取值。

定义 2　给定任一微簇 c_i 与空间内任意点 X,c_i 对 X 的近似密度吸引定义 $attr_{ap}(c_i,X)$ 如下:

$$attr_{ap}(c_i,X) = \frac{n(c_i)}{nh} K\left(\frac{X - c(c_i)}{h}\right) \tag{3}$$

基于这一概念,我们证明了如下定理:

定理 1　空间内任一点 X,其密度估计近似为所有微簇对其近似密度吸引之和,即:

$$\hat{f}(X) \approx \sum_{c_i \in M} attr_{ap}(c_i,X) \tag{4}$$

证明　略。

定理 1 表明,近似的相对误差与阈值 ε 和平滑参数 h 相关,当阈值 ε 足够小时,可以通过观察各个微簇的近似质心与数据点个数来检测一干预事件对空间内任意一点密度变化的影响。下面定义 3 中提出微簇的密度吸引特征向量融合了这两个变量。

定义 3 d 维数据空间内,给定任一微簇 c_i,其密度吸引特征向量定义为向量:

$$\boldsymbol{af}(c_i) = (c(c_i), n(c_i)) \tag{5}$$

非实时的微聚类过程中,用户给定干预时间变量 I_t^T 和干预分析时间段 $[T_1, T_2]$ $(T_1 \leqslant T \leqslant T_2)$,微聚类干预模型针对各个微簇的密度吸引特征向量,建立 ARIMA 回归干预模型,从而得到干预变量的回归模型:

$$z_t = \boldsymbol{v}(B)\boldsymbol{I}_t, \quad t \geqslant T \tag{6}$$

其中:

$$\begin{cases} \boldsymbol{I}_t = (I_t^T, \cdots, I_t^T) \\ \boldsymbol{v}(B) = (v(B)_1, v(B)_2, \cdots, v(B)_{d+1})^{\mathrm{T}} \end{cases} \tag{7}$$

其中 $v(B)_j (1 \leqslant j \leqslant d+1)$ 为回归模型的系数部分。

在实时的微聚类步骤中,我们通过类似 CluStream 的算法[21],首先调用 k-means 创建初始微聚类,然后随着流数据到达,将每个数据点归入离其最近的微簇,同时修改其簇特征 CF。本算法所需的微聚类的半径和质心都可以由簇特征 CF 在 $O(d)$ 算法内得到,d 为数据空间的维度。

我们在真实的 KDD-CUP 99 网络入侵检测数据集上做了若干实验。该数据集中标准网络入侵事件发生和结束的时间,我们以 smurf, back, neptune 三种不同的网络入侵作为其数据流的干预时间,进行干预挖掘。

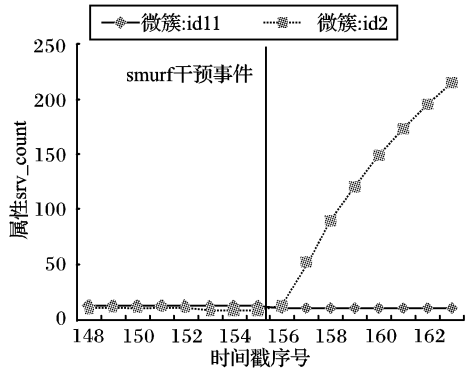

图 3 密度吸引特征向量在 smurf 干预下的变化

图 3 ~ 5 是挖掘得到的三个不同的干预事件回归模型,图 3 中,$c_{11}: z_t^{27} = 101.613 - 0.816712 z_{t-1}^{27} + 1.58961 z_{t-2}^{27}$, $c_2: z_t^{27} = -0.309481 - 0.383527 z_{t-1}^{27} + 1.24711 z_{t-2}^{27}$。图 4 中,$c_{57}: z^t 1 = 23043.6 + 0.478956 z_{t-1}^1 + 0.0926349 z_{t-2}^1$, $c_5: z^t 1 = -16.7531 - 0.369151 z_{t-1}^1 + 1.64515 z_{t-2}^1$。图 5 中,$c_0: z_t^{19} = -0.18058 - 0.29454 z_{t-1}^{19} + 1.43227 z_{t-2}^{19}$, $c_8: z_t^{19} = 18.5339 - 0.21562 z_{t-1}^{19} + 1.09343 z_{t-2}^{19}$。

由于篇幅有限,本文仅列举其中影响最大的微簇与另外一微簇在干预事件发生后的对比。图 3 显示的是微簇 id11 与 id2 在 smurf 入侵攻击下,两个微簇的密度吸引特征向量的属性“srv_count”上的干预变量回归模型。图 3 下方,z_t^{27} 代表了干预变量的第 27 维属性值,即属性“srv_count”。图 4、5 分别显示了另外两个微簇的密度吸引特征向量属性“src_bytes”与属性“count”值在 back 与 neptune 干预发生前后的变化。

通过 KDD-CUP 99 网络入侵数据的实验结果表明,smurf, back, neptune 三种不同的拒绝服务(Denial of Service, DoS)网络入侵主要影响的是高 srv_count, count 属性的微簇,这与我们对 DoS 网络入侵的先验知识是一致的。

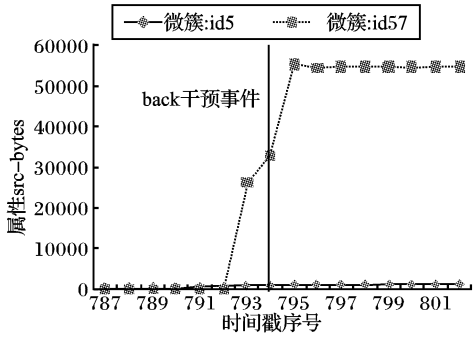

图 4 密度吸引特征向量在 back 干预下的变化

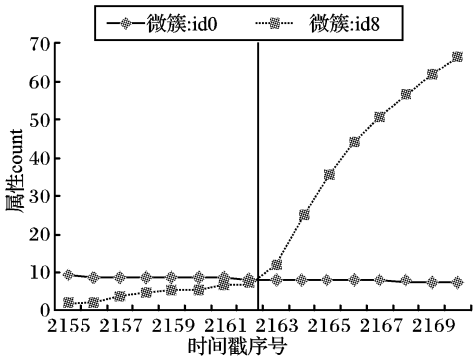

图 5 密度吸引特征向量在 neptune 干预下的变化

5 结语

本文综述亚复杂系统干预规则和干预动力学挖掘的研究背景,典型问题,相关概念和模型,介绍作者在朴素干预规则和数值型干预规则挖掘以及基于密度的数据流干预分析模型的初步结果。亚复杂系统干预规则挖掘是复杂的有挑战性的课题,有大量问题和猜想尚待研究,如干预分型和干预代数、干预的解释和评价等。

参考文献:

[1] 王安麟. 复杂系统的分析与建模[M]. 上海: 上海交通大学出版社, 2003.

[2] LUGER G F. 人工智能: 复杂问题求解的结构和策略[M]. 4 版. 北京: 机械工业出版社, 2003.

[3] BADII R, POLITI A. 复杂性: 物理学中的递阶结构和标度[M]. 北京: 清华大学出版社, 2000.

[4] 克劳斯. 迈因策尔. 复杂性中的思维[M]. 北京: 中央编译出版社, 2002.

[5] 代礼, 朱军, 周光萱, 等. 综合征性神经管缺陷 3789 例分析[J]. 中华妇产科杂志, 2003, 38(1): 17 - 19.

[6] 周光萱, 朱军, 代礼, 等. 1996 至 2000 年全国先天性腹裂畸形监测资料分析[J]. 中华预防医学杂志, 2005, 39(4): 257 - 259.

[7] 李科生, 蒲玮, 朱军. 计划生育技术人员有关增补叶酸预防胎儿神经管缺陷的 KAP 调查分析[J]. 中国妇幼保健, 2005, 20(8): 1008 - 1010.

[8] 李科生, 蒲玮, 朱军. 计划生育技术服务人员有关神经管缺陷及其干预措施知晓率的调查[J]. 中国妇幼保健, 2005, 20(9): 1133 - 1135.

[9] KHAN A, REVETT K. Data mining the PIMA dataset using rough set theory with a special emphasis on rule reduction[J]. Proceedings of the 8th International Multitopic Conference: INMIC 2004. Washington, DC: IEEE Computer Society, 2004: 334 - 339.

(下转第 2748 页)

统的整体性能在不断提高,未来将会有越来越多的自动问答系统投入实际应用中。

参考文献:

[1] VOORHEES E M, TICE D M. Building a question answering test collection[C]// Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2000: 200-207.

[2] TREC. The Text REtrieval Conference (TREC)[EB/OL]. [2008-02-15]. http://trec.nist.gov/.

[3] ZHANG D, LEE W S. Question classification using support vector machines[C]// Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2003:26-32.

[4] MOSCHITTI A. Efficient convolution kernels for dependency and constituent syntactic trees[C]// Proceedings of the 17th European Conference on Machine Learning, LNCS 4212. Berlin: Springer-Verlag, 2006:318-329.

[5] LI X, ROTH D. Learning question classifiers: The role of semantic information[J]. Journal of Natural Language Engineering, 2005, 12(3):229-249.

[6] BLUNSOM P, KOCIK K, CURRAN J R. Question classification with log-linear models[C]// Proceedings Of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2006: 615-616.

[7] TELLEX S, KATZ B, LIN J, *et al*. Quantitative evaluation of passage retrieval algorithms for question answering[C]// Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2003: 41-47.

[8] SUN R, ONG C H, CHUA T S. Mining dependency relations for query expansion in passage retrieval[C]// Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2006: 382-389.

[9] METZLER D, CROFT W B. Latent concept expansion using Markov random fields[C]// Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2007: 311-318.

[10] CUI H, SUN R, LI K, *et al*. Question answering passage retrieval using dependency relations[C]// Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2005: 400-407.

[11] LIN J, QUAN D, SINHA V, *et al*. What makes a good answer? The role of context in question answering[C]// Proceedings of the 9th IFIP TC13 International Conference on Human-Computer Interaction. Zurich, Switzerland: [s.n.], 2003:25-32.

[12] KAISSER M, BECKER T. Question answering by searching large corpora with linguistic methods[EB/OL]. [2008-03-016]. http://trec.nist.gov/pubs/trec13/papers/saarlandu.qa.pdf.

[13] SHEN D, KRUIJFF G J M, KLAKOW D. Exploring syntactic relation patterns for question answering[C]// Proceedings of the 2nd International Joint Conference on Natural Language, LNCS 3651. Berlin: Springer-Verlag, 2005: 507-518.

[14] HAN K.S, SONG Y I, RIM H C. Probabilistic model for definitional question answering[C]// Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2006: 212-219.

[15] KO J, SI L, NYBERG E. A probabilistic graphical model for joint answer ranking in question answering[C]// Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2007: 343-350.

(上接第2736页)

[10] GIARDINA M, HUO YONG-YANG, AZUAJE F, *et al*. A missing data estimation analysis in type II diabetes databases [C]// Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems: CBMS'05. Washington, DC: IEEE Computer Society, 2005: 347-352.

[11] HUANG Y, MCCULLAGH P, BLACK N, *et al*. Feature selection and classification model construction on type 2 diabetic patients' data[C]// Proceedings of the 4th Industrial Conference on Data Mining: ICDM 2004, LNCS 3275. Berlin: Springer-Verlag, 2004:153-162.

[12] 王恒.2型糖尿病发病危险因素与血糖值变化关系的研究[D].北京:北京理工大学,2006.

[13] BOX G E P, TIAO G C. Intervention analysis with applications to economic and environmental problems[J]. Journal of the American Statistical Association, 1975,(70):70-79.

[14] 李虹,蔡之华.关联规则在医疗数据分析中的应用[J].微机发展,2003,13(6):94-97.

[15] ORDONEZ C, SANTANA C A, de BRAAL L. Discovery interesting association rules in medical data [EB/OL]. [2008-02-18]. http://citeseer.nj.nec.com/ordonez00discovering.html.

[16] AGRAWAL R, SRIKANT R. Fast Algorithms for mining association rules in large databases[C]// Proceedings of the 20th International Conference on Very Large Data Bases. San Francisco, CA: Morgan Kaufmann Publishers, 1994:478-499.

[17] HAN J, PENG J, YIN Y, *et al*. Mining frequent patterns without candidate generation[C]// Proceedings of the 2000 ACM SIGMOD International Conference On Management of Data. New York: ACM Press, 2000:1-12.

[18] 段磊,唐常杰,左劼,等.基于基因表达式编程的抗噪声数据的函数挖掘方法[J].计算机研究与发展,2004,41(10):1684-1689.

[19] FERREIRA C. Gene expression programming: Mathematical modeling by an artificial intelligence[M]. 2nd ed. Berlin: Springer-Verlag, 2006.

[20] SILVERMAN B W. Density estimation for statistics and data analysis[M]. London: Chapman and Hall/CRC, 1986.

[21] AGGARWAL C C, HAN J, WANG J, *et al*. A framework for clustering evolving data streams[C]// Proceedings of the 29th International Conference on Very Large Data Bases. Berlin: VLDB Endowment, 2003:81-92.