文章编号:1001 - 9081(2010)01 - 0010 - 05

干预规则挖掘的任务分类和三项技术进展

唐常杰[1],段　磊[1],王　悦[1],杨　宁[1],朱　军[2],代　礼[2]
(1. 四川大学 计算机学院,成都 610065; 2. 中国出生缺陷监测中心,成都 610064)
(cjtang@scu.edu.cn; leiduan@scu.edu.cn)

摘　要:介绍了亚复杂系统中干预规则的基本概念和挖掘方法,提出了干预规则挖掘技术的分类准则,综述了三项干预规则挖掘技术的最新进展,包括疾病状态干预技术、基于数据流的未知干预发现技术和基于并行事件序列的干预规则挖掘。在实践基础上分析了干预规则挖掘的难点,展望了进一步的研究工作。

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

Task classification of intervention rules mining and advances of three technologies

TANG Chang-jie[1], DUAN Lei[1], WANG Yue[1], YANG Ning[1], ZHU Jun[2], DAI Li[2]
(1. *School of Computer Science, Sichuan University, Chengdu Sichuan* 610065, *China*;
2. *National Center for Birth Defects Monitoring, Chengdu Sichuan* 610064, *China*)

Abstract: The main contributions of this paper include: (1) introducing the basic concepts and mining methods of intervention rule over sub-complex system; (2) proposing the classification criteria for the tasks of intervention rules mining; (3) surveying the advances on three special mining techniques for intervention rules, including disease state intervention, intervention discovery from data streams, and intervention mining from parallel event sequences; (4) discussing the challenges and future research of intervention rules mining.

Key words: intervention rules mining; sub-complex system; data mining

0　引言

近年来,数据处理技术得到了广泛研究和深刻进步。从哲学上看,数据处理活动有 4 个层次,即搜集、存储、挖掘和干预。挖掘是为了认识自然,干预是在尊重自然的前提下改造自然,是数据处理活动的最高境界。典型干预实践包括局部气候干预、金融危机下经济调控、糖尿病干预,出生缺陷干预,等等。

干预不是藐视自然,而是“循律促变”;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。通常含 3 个要点:1)挖掘数据干预动力学规律,即干预规则;2)挖掘指定对象在指定定干预下的响应规律,发现最有效的干预措施和力度;3)见微知著,从对象的微变分析,发现和预测外界干预因素。分析结果广泛应用于工程、科研实践、社会调控(金融体系、国家政策评价)等领域,为决策提供有力的依据。

为简化研究对象、降低研究难度,本文提出了“亚复杂系统”模型[1-2]。它是在对复杂系统进行特征提取、抽取主要因素、降低数据维度等操作后,得到的目前有可能做出工程性解决方案的系统。具有下列优点:1)屏蔽了复杂系统中次要或噪声因素;2)便于“分而治之”地分析问题。目前干预规则挖掘主要在亚复杂系统上进行。

文献[3]提出了“what-if”分析。其要点是:1)在历史数据上建模;2)发现类似与“某种力度的某种行(某种后果”,的规则;3)假定这种规律在未来也成立,从而预测在未来某种干预措施实施后可能产生的影响。

1　干预规则挖掘任务分类

为简明地表达思想,本文给出必要的符号和术语,避免复杂的形式化描述,通过较多例子来说明思路。

被干预对象记为 O,其状态记为 S_p,干预手段 F,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:

$$Inv(O, F): S_p \xrightarrow{F} S_i$$

干预规则挖掘分为三类,下面举例说明。

1) RuleClass1。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。

2) RuleClass2。已知当前状态 S_p,干预措施 F,设 ε 描述干预措施的实施参数(如:强度、频率等),求达到预期干预效果 S_e,ε 的取值。

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?

收稿日期:2009 - 07 - 02;**修回日期**:2009 - 08 - 25。
基金项目:国家自然科学基金资助项目(60773169);国家“十一五”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; 段磊(1981 -),男,四川成都人,讲师,博士,主要研究方向:数据挖掘; 王悦(1981 -),男,四川成都人,博士研究生,主要研究方向:数据库与知识工程; 杨宁(1974 -),男,四川成都人,博士研究生,主要研究方向:数据库与知识发现; 朱军(1964 -),女,四川成都人,研究员,主要研究方向:出生缺陷干预; 代礼(1974 -),男,四川三台人,副教授,博士,主要研究方向:出生缺陷干预。

该类任务的另一种形式是求取要达到最佳干预效果,干预措施的实施参数。

3)RuleClass3。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 3　已知某地区某类出生缺陷发病率为 p。通过对比其他地域同类缺陷发病情况,利用数据挖掘方法挖掘能使该缺陷发病率降低 p' 的潜在规则。

上述三类任务中,挖掘能到达预期干预效果的干预规则难度最大。通常还需要结合领域知识。

根据数据对象的性质,可将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。

根据对象的分类基数,可将干预规则挖掘任务分为:

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:对某一地区产妇进行出生婴儿缺陷干预。

在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地应用于其他领域中。

2　疾病状态干预技术

基因疗法和药物干预[4-5]旨在将对象(如:病例样本、病变组织、病变器官)的状态从不良转变为良好,我们称此技术为疾病状态干预。即,对给定的病例样本集(包含不良状态和良好状态)挖掘可以将疾病状态从不良转变为良好的干预方法。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 说明了疾病状态干预问题。

例 4　一个 5- 基因(G_1, …, G_5)的病例样本,设 t_d = (0,1,1,0,0) 是一不良状态的样本,t_n = (1,1,1,0,1) 是一良好状态的样本。其中,样本中的 0、1 分别代表基因的基因表达水平为低或高。容易看出,若将样本 t_d 中的基因 G_1 和 G_5 的表达水平从低转变为高,样本 t_d 将从不良状态转换为良好状态。因此 G_1、G_5 称为状态转换项,组成的集合称为状态转换集。

例 4 表明,样本状态可以通过改变其属性而发生变化。为提高干预的实用性转换集包含的属性个数尽量少,并且属性改变的代价尽量小。作者实践了下列技术要点:

1)干预对象分类。在疾病状态干预问题中,我们应用分类模型来确立被干预对象是否处于良好状态。若 f 是一个分类器,对象 o 是处于不良状态,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 的目标即是 $f(X(o)) = des$。

2)融合多个分类模型。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,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。根据被干预对象是一个病例或一类病例,将疾病状态干预分为个体疾病状态干预和全体疾病状态干预。

3)挖掘状态转换项算法。有 3 个要素: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替代的候选状态转换项。

2.1　个体疾病状态干预挖掘方法

给定不良状态的病例样本 o_u,算法技术要点如下:

1)确定候选状态转换项。确定待改变的属性及其改变的目标值。对任意两个病例样本 o_1, o_2,我们根据它们之间不同属性值的个数确定其差异度:

$$dis(o_1, o_2) = |\{A \mid o_1(A) \neq o_2(A)\}|$$

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,记为 R。R 中每样本 o_u 的参考样本。

设 $difAtt(o_u, R) = \{A \mid o_u(A) \neq o_d(A), o_d \in R\}$,那么集合 $difAtt(o_u, R)$ 中包含的属性即是候选的待改变属性。此外,还需要确定每个待改变属性的目标值。对属性 A,设 $T = \{o(A) \mid o \in D_{des}, o(A) \neq o_u(A)\}$,即属性 A 的所有候选目标值。依条件概率选取最佳目标值 v_t:

$$Pr(state = des \mid A = v_t) = \max\{Pr(state = des \mid A = v) \mid v(T\}$$

其中,$state = des$ 表明样本状态为良好状态。这样,将 A 的值改变为 v_t 将有利于提高干预成功的概率。对所有 $difAtt(o_u, R)$ 中的属性及其改变目标值,记为 $CandCItems = \{(A, v_t) \mid A$ in $difAtt(o_u, R), (A, v_t)$ 是 A 的状态转换项$\}$。

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?②o_u 的参考样本中有多少样本在属性 A 上的取值与 o_u 不同?

对于①,希望属性 A 在样本 o_u 上的取值 $o_u(A)$ 与不良状态强相关,并且目标值与良好状态强相关。对这类属性进行干预能有效地达到样本状态转变的目的。对 o_u 的每个候选状态转换项 (A, v) 定义全局转换效益(Global Vonversion Utility, GCU):

$$GCU(A, o_u(A), v) = Pr(state = und \mid A = o_u(A)) * Pr(state = des \mid A = v)$$

尽管 GCU 只对一个不良状态的病例样本挖掘状态转换集,GCU 考虑了不良状态和良好状态中的所有样本。具有较高 GCU 表明该属性在大多数不良状态病例中表现为当前值,而在大多良好状态中表现为改变目标值。

对于②,我们根据参考样本集 (R) 考察属性的局部转换效益(Reference Conversion Utility,RCU):

$$RCU(A, o_u(A)) = |\{o \in R \mid o(A) \neq o_u(A)\}| / |R|$$

较高 RCU 值的属性表明属性的改变目标值出现较少。

为成功转换病例状态,我们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。这样,依据 GCU 和 RCU,定义属性 A 的转换效益(Conversion Utility,CU)如下:

$$CU(A, o_u(A), v) = GCU(A, o_u(A), v) * RCU(A, o_u(A))$$

当病例样本有多个参考样本时,RCU 在 CU 计算中具有较大影响。当只有一个参考样本时,RCU 为常量 1,对 CU 计算没有影响。可见,在 CU 计算中考虑 RCU 和 GCU 能适用于处理任意情况。

倘若状态转换集的大小为 s。一种办法是选取 s 个具有最大 CU 值的状态转换项。但这样没有考虑不同状态转换项之间的关联。为了达到最佳的干预效果,应当尽量选取相关性小的状态干预项组成状态干预集。

在病例状态干预问题中,我们根据属性的取值分布衡量属性之间的相关性。设 $x = (A_x, v_x)$ 和 $y = (A_y, v_y)$ 是两个状态干预项,D_{des} 是良好状态病例集,那么它们之间的关联度

(COR)定义如下:

$$COR(x, y) = 1 - count_{des}(A_x = v_x, A_y = v_y) / \max\{count_{des}(A_x = v_x), count_{des}(A_y = v_y)\} + 1/|D_{des}|$$

其中,$count_{des}(A_x = v_x)$ 是 D_{des} 中 $o(A_x) = v_x$ 的样本个数。增加 $1/|D_{des}|$ 来保证 $COR(x, y) > 0$。容易看出,如果干预项 x 和 y 有很多共同出现的元组,那么它们的相关性强。COR 的取值范围为 $[1/|D_{des}|, 1 + 1/|D_{des}|]$,值越大的 $COR(x, y)$ 表明干预项之间的关联性越小。

由于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。

2.2 类型疾病状态干预

旨在对所有处于不良状态的病例样本挖掘状态干预集实现干预。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平凡且相关性低的状态干预项组成全体疾病状态干预集。

表 1 是真实数据集上的实验结果。采用 Logistic、SMO、ID3 (J48)、Naive Bayes、RBFNetwork、KNN (K = 3)、Hyper Pipes (HP) 和 Voting Feature Intervals (VFI)对干预结果分类,并记录干预成功样本占总样本比率的平均值。

表 1 实验数据集

Dataset	# samples in D_{des}	# samples in D_{und}	# attr.
Breast Cancer	44 (non-relapse)	34 (relapse)	619
Colon Cancer	22 (normal)	40 (cancer)	135
Leukemia	11 (AML)	27 (ALL)	865
Lung Cancer	16 (Mesothelioma)	16 (ADCA)	2 [illegible]3
Prostate Cancer	50 (normal)	52 (cancer)	1 [illegible]34
Mushroom	4208 (edible)	3 916 (poisonous)	22
Tic-Tac-Toe	626 (positive)	332 (negative)	9
Congressional Voting Records	168 (democrat)	267 (republican)	16

作者提出的方法记为 DSCMiner。实验中对比了 WEKA[6] 提供的特征评价算法 ReliefFAttributeEval 和 SVMAttributeEval,以及两个特征选择算法 FCBF[7] 和 INTERACT[8]。根据这些算法的属性评价,组成状态干预集,分别记为 ReliefF-AE,,SVM-AE,FCBF-AE 和 INTERACT-AE。图 1 比较了干预成功率,表明 DSCMiner 方法在大多数数据集中都能取得最好的干预效果。

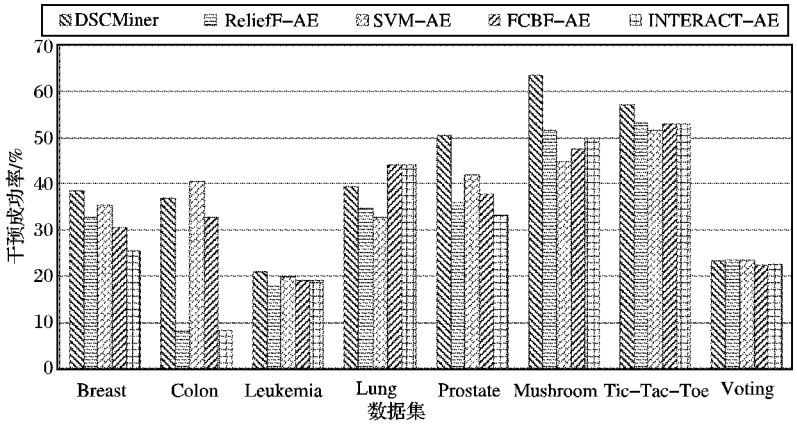


图 1 干预成功率比较

3 基于数据流的未知干预发现技术

出生缺陷干预实践中,未知的外部干预有重要意义,为研究缺失但非常重要的干预事件,本文提出了干预事件规划模型(Intervention Events Scheduler,IES)。IES 在传统数据流异常挖掘方法[9-10]上进行发展,思路是根据观察数据的统计特性变化来推断未知干预事件的存在。例 5 说明了干预事件规划模型的研究动机。

例 5 考虑在高速公路上的交通情况预测问题:洛杉矶 101-北高速公路的一个斜坡上安装了一系列循环监控传感器,监控体育馆周围的交通流。通过交通流变化趋势,自适应地检测体育馆中比赛事件的发生(如图 2)。

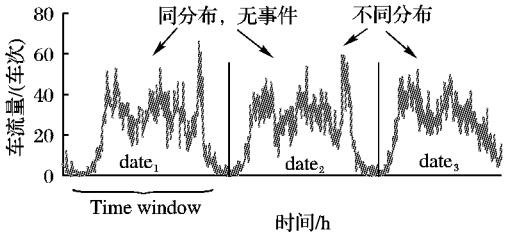


图 2 由传感器搜集到的交通流数据

观察 1 图 2 中交通流数据随时间变化,时间窗口中数据分布满足一定规律。图中 $date_1$ 与 $date_2$ 时窗中数据分布相似;而 $date_2$ 与 $date_3$ 却有明显的数据分布差距,提示有未知干预(如赛事)发生。

为分析数据流干预,IES 借鉴了热力学中配容熵(表示为 C_E,计算方法见式(1))来描述这一观察:当两个分布相近时它们的配容熵值 C_E 接近;反之,配容熵的差距提示有未知干预事件发生。

$$C_E(t) = \log\left(N! / \prod_{i=1}^{k}(N_i!)\right) \quad (1)$$

根据这一朴素观察,提出下列假设:

稳定性假设 在数据流环境下,如无外界干预,观察对象的分布将会相对稳定;亦即数据流状态变化提示有干预事件发生。

本文提出了 IES 解决了下列个问题:1)根据数据流变化自适应发现干预事件;2)分析干预事件影响力;3)预测未来干预事件的发生。IES 结构如下:

定义 1 干预规划模型 IES 是一个四元组 $(O(t), Q, P, C, S(t))$,其中:

1) $O(t) = \{item_1, item_2, item_3, \cdots, item_t\}$ 是一个观察序列,t 是观察时间点;

2) $Q = \{q_1, q_2, \cdots, q_n\}$ 为所有干预事件的有限集合;

3) M_t 是干预事件之间的概率转移矩阵;

4) M_C 是干预事件到观察数据间的混淆概率矩阵;

5) $S(t) = \{event_1, event_2, event_3, \cdots, event_t\}$ 为干预事的序列,t 为时间点;

IES 模型含一条观察数据流,一条干预事件流,状态转移矩阵、混淆矩阵。并有事件检测、相关的矩阵创建以及干预预测算法。干预检测算法将配容熵、离均差组合起来,使用序列显著差 $P_s(t)$ 来监测数据流的变化情况。根据情况设相关阈值 k 来检验异常情况的发生($P_s(t) > k$ 表明未知干预事件发生)。序列显著差 $P_s(t)$ 计算方法如下:

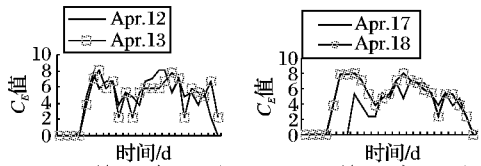
$$N(t) = SquareSum(S_{CE}(t), S_{CE}(t+1)) \quad (2)$$

$$Inf(t) = \begin{cases} N(t)/N(t-1), & N(t) > N(t-1) \\ N(t-1)/N(t), & N(t-1) \leqslant N(t) \end{cases} \quad (3)$$

$$P_s(t) = Inf(t) \times Inf(t+1) \quad (4)$$

IES 模型遵从隐马尔可夫特性[11]:将观察到的数据作为外在表现,而事件发生与否的状态作为隐藏特性,建立从状态到外部数据的混淆矩阵(Confusion Matrix),状态之间的转移矩阵。

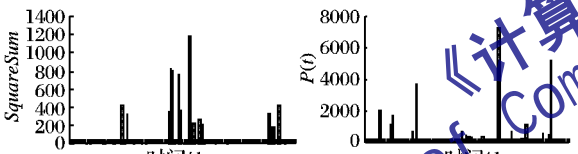
在真实的交通流数据上的实验,文献[12]验证了 IES 模型和关算法的正确性以及完备性。图 3 展示了两个交通流的 C_E 序列:经过对原始观察数据的转换,可清楚地区别出两段相似(图 3a)与不同(图 3b)的序列。提示(a)在[2005-4-12, 2005-4-13]无干预事件发生;(b)在[2005-4-17, 2005-4-18]明显有未知干预事件发生。



(a) 4月12日与13日之间的C_E值比较　(b) 4月17日与18日之间的C_E值比较

图 3　C_E 值比较实验

对两个交通流的 C_E 序列计算离均差,最后获取到序列显著差。通过每段时刻的序列显著差与阈值的差就能检测出干预事件。在文献[12]对全部数据[2005-4-12, 2005-10-1]进行计算,结果见图 4。



(a) 离均差结果　(b) 序列显著差结果

图 4　序列显著差计算

根据图 4(b)的数据表明,当阈值 k 取 30 时,检测正确率为65.4%;当 k 取12时,正确率为92.6%;当 k 被设置为11时,全部81个有记录的比赛事件均从交通流数据中检测到。文献[12]对干预预测算法进行了测试,使用了有记录的 81 个事件中的一半与相关时间的交通流作为训练数据训练 IES 模型。整个训练过程消耗 550 ms。最后根据获取的模型使用 intervention-prediction 算法增量预测可能发生的干预事件,其正确率为 61.4%。

4　基于并行事件序列的干预规则挖掘

在多个事件序列之间进行相关性分析已经受到广泛的关注,但是这种相关性无法解释因果问题。我们在实践中发现并行事件序列之间普遍存在着相互干预,提出了新的“干预时态模式”,并研究了相应挖掘算法。

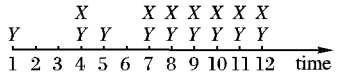


图 5　由两种类型构成的并行事件序列片段

例 6　图 5 中,X 和 Y 是两个不同类型的事件序列 $X(t)$ 和 $Y(t)$,t 是时间且 $1 \leqslant t \leqslant 12$。设观察周期为 2,则整个持续时间可以划分为 6 个周期。在第 i 个周期内 X、Y 出现的次数分别记为 x_i 和 y_i,$1 \leqslant i \leqslant 6$。图 5 表明,每当 X 的新事件出现,Y 的新事件增多,即 y_{i+1} 受到了 x_i 的影响,我们称此时存在从 X 到 Y 的干预,记为 $X \to Y$。如果每个事件序列是一个 Markov 链,则上述干预现象在本质上可以解释为对广义 Markov 性质的背离。根据广义 Markov 性质,$P(y_{i+1} \mid y_i) = P(y_{i+1} \mid y_i, x_i)$,当存在干预时,$P(y_{i+1} \mid y_i) \neq P(y_{i+1} \mid y_i, x_i)$。这种背离的程度可以用 Kullback-Leibler 散度衡量。

将上述观察和思考形式化,引入下列概念。

定义 2　设 $X(t)$ 和 $Y(t)$ 是两个事件序列,$P(\cdot)$ 是概率函数,如果 $P(y_{i+1} \mid y_i) \neq P(y_{i+1} \mid y_i, x_i)$ 成立,则称 X 干预 Y,记为 $X \to Y$,称 X 是干预源,Y 是干预目标。

定义 3　设 $X(t)$ 和 $Y(t)$ 是两个事件序列,$X \to Y$ 的干预强度定义为 $I(X \to Y) = \sum P(y_{i+1}, y_i, x_i) \log P(y_{i+1} \mid y_i, x_i)/P(y_{i+1} \mid y_i)$。

干预强度中出现的联合概率的计算可以通过核密度估计完成. 给定一个解析度 r, $P(y_{i+1}, y_i, x_i)$ 计算如下:$P(y_{i+1}, y_i, x_i) = (1/N)\sum_j K(\max(|y_{i+1} - y_{j+1}|, |y_i - y_j|, |x_i - x_j|) - r)$,其中 $K(\cdot)$ 满足 $K(x > 0) = 1, K(x \leqslant 0) = 0$,$N$ 是 ij 对的数量。类似的,$P(y_{i+1}, y_i)$,$P(y_i, x_i)$ 和 $P(y_i)$ 分别计算如下:1) $P(y_{i+1}, y_i) = (1/N)\sum_j K(\max(|y_{i+1} - y_{j+1}|, |y_i - y_j|) - r)$;2) $P(y_i, x_i) = (1/N)\sum_j K(\max(|y_i - y_j|, |x_i - x_j|) - r)$;3) $P(y_i) = (1/N)\sum_j K(|y_i - y_j| - r)$。于是条件概率可以计算如下:$P(y_{i+1} \mid y_i, x_i) = P(y_{i+1}, y_i, x_i)/P(y_i, x_i)$,$P(y_{i+1} \mid y_i) = P(y_{i+1}, y_i)/P(y_i)$。

在文献[13]中,根据 Markov 链的性质和 Kullback-Leibler 散度的单向性,证明了干预和干预强度具有下述性质:

命题 1　$X \to Y$ 是自反的、反对称的、传递的。

命题 2　$I(X \to Y) \neq I(Y \to X)$。

如果不存在噪声,则当 $X \to Y$ 成立时,$I(X \to Y) > 0$,$I(Y(X) = 0$。但是现实数据中一般充满噪声,$I(Y \to X) = 0$ 不一定成立。由于命题 2 保证了 $I(X \to Y) \neq I(Y \to X)$,因此我们可以通过比较 $I(X \to Y)$ 和 $I(Y \to X)$ 大小来判断 X 和 Y 之间的干预的方向,当 $I(X \to Y) > I(Y \to X)$ 时,$X \to Y$ 成立,反之,$Y \to X$ 成立。

为揭示并行事件序列之间干预的本质,从熵率观点进行分析。根据信息论,随机过程 $Y(t)$ 熵率定义为 $h_y = \lim_{n \to \infty}(1/n)H(y_1, y_2, \cdots, y_n)$,当 $Y(t)$ 是 Markov 过程时,$h_y = H(y_{i+1} \mid y_i)$,其中 $H(\cdot)$ 是信息熵。我们证明了下述命题[13]:

命题 3　设 $X(t)$ 和 $Y(t)$ 是两个事件序,$I(X \to Y) = h_y - h_{y|x}$,其中 $h_y = H(y_{i+1} \mid y_i)$,$h_{y|x} = H(y_{i+1} \mid y_i, x_i)$。

命题 3 表明:1) 当没有干预时,事件序列 $Y(t)$ 内在的不确定性(即信息熵)将以 $h_y = H(y_{i+1} \mid y_i)$ 的速率增长;2) 当 $X(t)$ 对 $Y(t)$ 实施干预后,$Y(t)$ 的不确定性增长速率将降低为 $h_{y|x} = H(y_{i+1} \mid y_i, x_i)$(注意 $H(y_{i+1} \mid y_i) \geqslant H(y_{i+1} \mid y_i, x_i)$),降低的量正好等于干预强度 $I(X \to Y)$。命题 3 揭示了干预演化实质,即干预降低了目标事件序列演化的熵率。

基于上述研究,文献[13]设计了在并行事件序列中挖掘干预的算法(Mine Interventions from Parallel Event Sequences, MIPES),并在下载自 UCI 的真实数据集 CalIT2 上进行了验证。CalIT2 记录了关于某栋建筑的两类事件:In-event(进入事件)和 Out-event(离开事件),每隔 30 min 记录一次,共记录了 3 个月。不失一般性,本文随机抽取了 7 月 24 日的数据进行分析,其原始分布如图 6 所示。

图 7 给出了解析度 r 对于干预强度计算的影响。

图 7 表明:1)数据存在噪声,所以 I(In-event→Out-event)和 I(Out-event→In-event)都不为零。2)因为 I(In-event→Out-event) > I(Out-event→In-event),故可确认 In-event→Out-event 成立,这一结果符合常识:进入大楼的人越多(少),出来的人越多(少)。3)随着 r 的增大,I(In-event→Out-event)和 I(Out-event→In-event)趋于相等,这说明,较小的解析度有助于隔离噪声。4)过小的 r(r < 3)会导致可用数据数量较少,造成输出的干预强度较小。

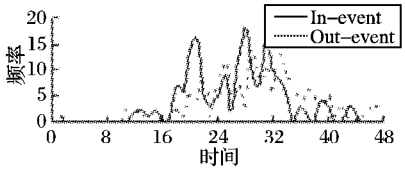


图 6 In-event 和 Out-event 的频率分布

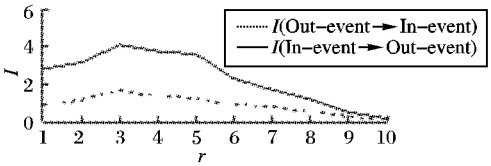


图 7 不同解析度时的干预强度

此外,实验还考察了算法 MIPES 的可伸缩性。图 8 表明,运行时间基本上随数据规模呈线性增长。因算法 MIPES 的时间复杂度主要依赖于事件类型的数量,而不是事件的数量,而在实践中,前者远远小于后者。

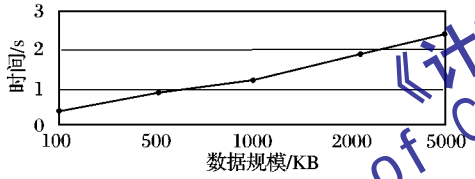


图 8 不同数据规模时算法 MIPES 的性能

5 结语

干预规则挖掘融合了多个数据挖掘研究方向,包括传统的关联、分类、时间序列等。在前述了三项技术研究实践中,作者深深体会到,干预规则挖掘研究还处于起步阶段。现阶段的难点,也是未来工作重点,包括:

1)被干预对象受外界影响大。干预行为类似化学反应过程,需要时间。在期待的响应回馈之前,被干预对象可能受外界其他影响,给观察、评估干预效果带来困扰。

2)对象数据不确定性。干预实施是较长过程,如我国对出生缺陷的检测已持续了 20 余年。长过程中数据可能受损,失真,产生了数据不确定性。目前,不确定性数据挖掘已引起研究者关注[14]。本文作者已开始了不确定环境下的干预规则挖掘,将在后续工作中发表。

3)高效亚复杂系统难以界定。亚复杂系统屏蔽复杂系统中无关因素以降低难度。从复杂系统中分离出高效、无损、低复杂度的亚复杂系统是挑战性问题。传统的特征提取、属性融合等方法有待提升以满足新需求。

4)单类训练数据。数据挖掘任务中训练数据通常有两类,但在干预分析中往往只有一类数据。因此,在干预效果验证、预测时,难以用传统对比挖掘方法评价干预效果,无法进一步建立干预模型,而且这也不同于 One-Class 问题[15]。一种思路是采用"历史 + 合理的假设"的方法来解决,目前正在探索中。

5)干预规则正确性验证。真实世界的复杂性远大于数据挖掘建立的模型。验证干预规则在真实世界的正确性需要合理、有效的方法。真实世界中,个体可能存在不可忽略的差异。为评价干预规则有效性带来了挑战。

6)领域知识的有效融合。领域知识能描述领域特殊的启发性知识。引导挖掘向正确方向进行。例如,某类出生缺陷同婴儿的性别相关,但领域知识会避免计算机找出改变性别这样的干预规则。有效地融合领域知识于干预规则挖掘,是一件有意义的工作。

参考文献:

[1] 唐常杰, 张悦, 唐良, 等. 亚复杂系统中动力学干预规则挖掘技术研究进展[J]. 计算机应用, 2008, 28(11):2732 - 2736.

[2] 张悦, 唐常杰, 李川, 等. 出生缺陷监测数据中的朴素干预规则挖掘[J]. 计算机科学与探索, 2009, 3(2):188 - 197.

[3] LAKSHMANAN V S, RUSSAKOVSKY A, SASHIKANTH V. What-if OLAP queries with changing dimensions[C]// Proceedings of the 2008 IEEE 24th International Conference on Data Engineering. Washington, DC: IEEE Computer Society, 2008:1334 - 1336.

[4] Gene Therapy[EB/OL]. [2009 - 04 - 15]. http://www.ornl.gov/sci/techresources/Human_Genome/medicine/genetherapy.shtml.

[5] Gene Therapy for Cancer[EB/OL]. [2009 - 04 - 20]. http://www.csupomona.edu/~isge/Gene Therapy'04.ppt.

[6] WITTEN I H, FRANK E. Data mining: Practical machine learning tools and techniques [M]. 2nd ed. San Francisco : Morgan Kaufmann, 2005.

[7] YU LEI, LIU HUAN. Feature selection for high-dimensional data: A fast correlation-based filter solution[EB/OL]. [2009 - 04 - 20]. http://www.hpl.hp.com/conferences/icml2003/papers/144.pdf.

[8] ZHAO ZHENG, LIU HUAN. Searching for interacting features [EB/OL]. [2009 - 04 - 20]. http://www.public.asu.edu/~huanliu/papers/ijcai07.pdf.

[9] CAI Y D, CLUSTTER D, PAPE G, *et al*. MAIDS: Mining alarming incidents from data streams[EB/OL] . [2009 - 04 - 25]. http://algdocs.ncsa.uiuc.edu/PB-20040613-1.pdf.

[10] CURRY C, GROSSMAN R, LOCKIE D, *et al*. Detecting changes in large data sets of payment card data: A case study[C]// Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery And Data Mining. New York: ACM, 2007:1018 - 1022.

[11] IHLER A, HUTCHINS J, SMYTH P. Adaptive event detection with time-varying poisson process[C]// Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York: ACM, 2006:207 - 216.

[12] WANG YUE, TANG CHANGJIE, LI CHUAN, *et al*. Intervention events detection and prediction in data streams[C]// Proceedings of the Joint International Conferences on Advances in Data and Web Management. Berlin: Springer-Verlag, 2009:519 - 525.

[13] YANG NING, TANG CHANGJIE, WANG YUE. Mining interventions from parallel event sequences[C]// Proceedings of the Joint International Conferences on Advances in Data and Web Management. Berlin: Springer-Verlag, 2009:297 - 307.

[14] 周傲英, 金澈清, 王国仁, 等. 不确定性数据管理技术研究综述[J]. 计算机学报, 2009, 32(1):1 - 16.

[15] MANEVITZ L M, YOUSEF M. One-class SVMs for document classification[J]. The Journal of Machine Learning Research, 2002, 2:139 - 154.