软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2011,22(2):285−297 [doi: 10.3724/SP.J.1001.2011.03829]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

在不确定数据集上挖掘优化的概率干预策略*

王 悦[1,2,3], 唐常杰[1+], 杨 宁[1], 张 悦[1], 李红军[1], 郑皎凌[1], 朱 军[4]

[1](四川大学 计算机学院,四川 成都 610065)

[2](高可信软件技术教育部重点实验室(北京大学),北京 100871)

[3](北京大学 信息科学技术学院,北京 100871)

[4](四川大学 华西医学院 中国出生缺陷监测中心,四川 成都 610065)

Mining Optimized Probabilistic Intervention Strategy over Uncertain Data Set

WANG Yue[1,2,3], TANG Chang-Jie[1+], YANG Ning[1], ZHANG Yue[1], LI Hong-Jun[1], ZHENG Jiao-Ling[1], ZHU Jun[4]

[1](College of Computer Science, Sichuan University, Chengdu 610065, China)

[2](Key Laboratory of High Confidence Software Technologies (Peking University), Ministry of Education, Beijing 100871, China)

[3](School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China)

[4](Birth Defects Supervising Center, Western China Medical School, Sichuan University, Chengdu, 610065)

+Corresponding author: E-mail: cjtang@scu.edu.cn

Wang Y, Tang CJ, Yang N, Zhang Y, Li HJ, Zheng JL, Zhu J. Mining optimized probabilistic intervention strategy over uncertain data set. *Journal of Software*, 2011,22(2):285−297. http://www.jos.org.cn/1000-9825/3829.htm

Abstract: This study provides models to analyze the intervention over uncertain data. The main contributions include: (1) It establishes a base model named Uncertain Surveillance; (2) It proposes probabilistic intervention strategies over uncertain surveillance with evaluating algorithms; (3) It gives extensive experiments and makes comparisons to show that the proposed model is highly precise and efficient in three order magnitudes of naïve methods.
Key words: uncertain data; probabilistic intervention strategy; strategy evaluate; data mining; mass data analysis

摘 要: 提出了不确定干预分析模型,主要工作包括:(1) 建立了用于多维不确定数据分析的不确定监测点模型(uncertain surveillance);(2) 建立了基于不确定监测点的不确定干预策略及挖掘评价算法;(3) 在真实数据及仿真数据上对所提出的两种算法作了大量实验比较,验证了所提出的干预策略评价优化算法具有较高精度,效率比朴素方法高出 3 个数量级,适合在实际系统中处理海量干预评价.
关键词: 不确定数据;概率干预策略;策略评价;数据挖掘;海量数据分析
中图法分类号: TP311 **文献标识码**: A

人类在科研实践中已经积累了海量的数据.由于采集设备、方法、手段和环境的差异,现有的数据中存在不确定性,数据可靠性及可信度受各种因素的影响.高效而准确地分析不确定数据对研究和实践具有重要意义.传统数据挖掘模型假定规律潜藏在训练数据中,对数据要求苛刻,需要严格的预处理,以去掉数据中的不确定

* 基金项目: 国家自然科学基金(600773169); 国家科技支撑计划(2006BAI05A01, 2009BAK63B08); 国家高技术研究发展计划(863)(2009AA01Z150)

收稿时间: 2009-07-06; 修改时间: 2009-10-09; 定稿时间: 2010-03-04

性.常用的预处理包括均值替代、基于 k 近邻的缺失值处理、基于空间自回归模型的插补等.其不足有:(1) 失真度较大,可能牺牲数据的部分原始特性,无法精确地反映原始数据;(2) 通用性较小,一种数据预处理技术往往针对特定形式的数据;(3) 协调性较差,当样本数据间存在偏差时,传统预处理可能得出自相矛盾的结果.

数据挖掘的理论认为[1],数据干预知识挖掘处理活动应经历 4 个不同的阶段:搜集、存储、理解(挖掘)和干预(决策活动).前两个阶段是准备,第 3 阶段是认识自然,第 4 阶段是在尊重自然的前提下改造自然.对天气的干预、医学对糖尿病的干预,以及用系列人工地震释放应力,实现对地震的干预,都是干预的典型应用.干预规则挖掘和评价是数据挖掘的新分支.旨在分析外界因素对观察对象的影响,根据对象变化来预测未知干预因素,广泛应用于工程、科研、国家政策评价等领域,为决策提供有力的依据.传统干预分析常采用时序分析描述在时间维度上的变化趋势,如序列间关系、自相关等.其缺点是,时序分析不能分析数据在其他维度(如空间维度、频域维度等)上的变化趋势,分析时必须在真实的干预手段上进行,无法作出前瞻性推理分析.

与已有方法不同,本文采用"假设策略→真实历史数据→得出结论"(类似于,又不同于 What-IF[2])的方式来挖掘真实历史数据.希望从真实历史数据(带有不确定性的)中得到一些假设干预策略的可行度分析,为策略、决策的制定提供更有力的量化依据.由于直接从全部原始数据中获取干预策略信息,进一步地克服了失真度大的问题;同时,本文的方法不需要数据先验分布前提,由此对各类数据上的应用具有一定的通用性.

研究动机.本文以国家科技支撑计划"中国出生缺陷干预评价"项目为研究背景.从 1986 年开始,国家投入巨资从卫生部下属的各级医院收集婴儿相关数据.决策部门希望得到出生缺陷的变化趋势,并评价政府对出生缺陷干预的效果.20 多年中,数据收集手段和数据库系统的变化和观测和记录手段差异,引入了潜在的数据不确定性.

在不确定数据上分析评价干预策略的效果本身也具有不确定性.如,政府在 1987 年 4 月期间投入 m 万人民币资金为偏远地区生育期孕妇补充叶酸等所需药品,评价的效果为 C.然而,由于外部环境的改变,若在另外一个时间段(如 2009 年 1 月)在该项上投入 m 万人民币,效果却不一定能达到 C.

本文旨在解决以下问题:(a) 在不确定数据上,计算的指定策略可能会有概率 p_0 获得效果 C,且计算出概率 p_0;(b) 解决在实际系统中,大量用户提交上海量干预策略评价时的处理性能问题.

面临的挑战.不确定数据的干预策略评价问题主要面临下面两个挑战:(1) 正确的模型.建立干预策略评价模型,统一反映真实的或假设的干预策略及干预后的效果,给出概率性干预效果的评估,给出某种干预效果的概率,为决策提供量化的支持.(2) 计算量大.重要干预策略常常体现为国家政策,可能涉及几十亿元投资和成亿人口的利益,常要求在海量数据上评价大量的干预策略,因而对计算资源需求大.

1 相关工作

1.1 不确定数据

常用不确定数据处理手段包括两类,即增加概率维度的方法和不增加概率维度的方法.

(1) 增加概率维度的方法.要求数据本身有潜在概率,即具有概率维度,或者能够通过一定的手段还原数据的概率维度,这种有潜在概率的数据称为不确定型概率数据.

不确定型概率数据可有多种不同的表现形式[3]:关系型、半结构化、流数据、移动对象数据等.文献[4]整理了不确定数据管理的相关工作;Cavallo 等人[5]为处理数据中的不确定性,提出了概率数据库(probabilistic database)理论;Fuhr 等人[6]在概率数据库的基础上解决其关系代数、数据整合及查询分析;在概率数据库理论的基础上,Chau 等人[7]正式提出了不确定数据挖掘,用于处理结构不确定与具有存在概率维度的数据.Green 等人[8]为不完整的及概率性数据建立了针对关系型不确定数据的可能性模型,证明了可能性世界的完备性等定律.Nierman 等人[4]提出了 p-文档模型,将概率附加于文档树上,根据子图的概率来管理不确定半结构化数据数据;文献[9,10]扩展了 p-文档模型,提出了概率树模型,以增强模型对不确定数据的表达能力.

前述的不确定模型存在若干不足,包括:(a) 组合爆炸问题.在海量数据规模中,不同组合的概率值过小,可能世界中的数据实例组合失去了其意义,例如,所有组合可能小于概率限制.(b) 维度引入问题.需要特定方法为数

据引入概率维度.由于目前的不确定数据大多是从确定性数据源处理所得到的,需要采用统计方式或经验办法加入概率维度.主观因素和随机错误往往引入新的不确定因素,造成分析精度下降.

(2) 不增加概率维度的方法.直接采用不确定分析建模较为客观.Tao 等人[11]在概率密度函数(probability density function,简称 PDF)使用范围相关查询及概率阈值查询(probabilistic threshold queries)的基础上将概率范围查询扩展到多维不确定数据,使其算法能够适应任意概率密度函数,提高了查询灵活性和算法效率;Pei 等人[12]在不确定数据上分析概率型 Skyline,采用不确定分析建模,按给定概率搜索 NBA 球星数据集里进攻型、防守型及综合型球员 Skyline.

1.2 干预知识挖掘

目前干预分析技术主要有两大方向,即来自统计学的 ARIMA,以及采用数据挖掘领域算法的分析.

统计学中采用 ARIMA 模型[13,14]来处理干预预测问题,BOX 等人在文献[15]中对方法作了总结,将 ARIMA 模型应用到经济、环境时序数据干预分析中;熊焰等人[16]对 ARIMA 模型进行扩展,将其应用到股票及工业化统计数据上,取得了较好的效果.ARIMA 模型识别步骤需要人工从原始数据中得出结论,缺乏客观性,在海量数据处理方面不能通用.近年来,Zhu 等人[17]扩展指数分布来检查邮件流中的突发事件;Ihler 等人[18]根据泊松分布假设提出自适应检测方法来检测车流量掘算法.概念漂移等方法[19]用于检测数据中高阶模式变化.离群点检测[20,21]是数据挖掘中的异常检测方法.这些方法广泛用于干预分析,监视观察数据整体状态变化.

我们曾结合 HMM 和流数据挖掘相关的成果提出了自适应干预事件检测[22],建立了朴素模型评价预测干预手段.唐常杰等人[2]提出了朴素干预规则理论,在事务数据库关联规则的基础上引入更高阶的知识级别,取得了不错的知识发现成果.其基本思路是根据事务数据库关联规则来生成干预规则.需要多次的数据库扫描,消耗资源较大,难以应用到更一般的数据流处理环境.What-If 分析[23]是一种根据已知历史数据发现未知结果可能性的分析思路,有较好的应用[24].由于多数 What-If 分析是基于应用构建,要使用 What-If 的分析思路,还需要重新仔细设计分析模型.另外,其运算性能及效率尚有很大提升空间.

与已有方法不同,本文采用直接建模方法,并融合干预挖掘和 What-If 分析的思路,建立不确定数据环境下的干预评价算法.本文相应地对算法处理效率进行优化,以使其满足海量不确定数据分析的需求.

2 概率干预评价模型及问题定义

本文基本思路是,用谓词逻辑表述干预策略,计算历史数据中该谓词的数据统计指标,根据该指标来探索干预策略可能的效果.拟解决问题:“假如使用干预策略 $strategy_1$,可能会对结果有什么影响”.能量化分析干预策略的效果是:(1) 给定干预策略 s_1 的强度、方式,推算未来能够达到某确定效果的概率 p;(2) 按照指定概率 p,找出所有能够以概率 p 达到确定效果干预策略的集合 S.为了准确地描述本文概念,引入表 1 中的基本符号和术语.

Table 1 A summary of frequently used notions

表 1 常用符号表

Symbol	Meaning
U	Uncertain surveillance
$a\{A==c\}$	Intervention predication
F	Intervention predication field
S	Intervention strategy
W	Intervention measure space
T	Time points set
$Y_{s:t}$	Intensity of s at time t
$\theta_{s:t}$	Impact of s at time t

2.1 不确定对象分析基础

实践表明,不确定性的来源主要有两个层次:(1) 元数据级的不确定性,如字段名、字段类型和长度可能会有错误;(2) 记录级的不确定性,可能存在由于记录错误造成的“假”数据.本文先考虑数据记录级不确定性,将元数

据级不确定性留作下一步工作.处理数据集时,重点关注对象的统计信息,如出现次数、最大值等.由于实例存在级的不确定性,对象的某项统计信息在指定维度上呈不确定变化,这种对象称为不确定监测点.

定义 1(不确定监测点(uncertain surveillance)). 设 D 为不确定数据集,T 为时间间隔集合,$T=\{t_1,t_2,t_3,\ldots,t_n\}$.对 $\forall t_i \in T(1\leqslant i\leqslant n)$, 监测点 $U(t_i)$ 为 t_i 上满足一组属性集合 $A(A=\{a_1,a_2,a_3,\ldots,a_n\})$ 的数据的集合,记为 $U(t_i)=\{d|\forall d.a_i\in A,d.time\_interval=t_i,d\in D\}$.若在 T 的时间间隔中,$|U(t_i)|(1\leqslant i\leqslant n)$ 的数值不总是相等,则称 U 为数据集 D 上,以 T 为不确定维度的不确定监测点:

例 1:对于出生缺陷医学数据,一个不确定监测点的实例为 U_1,其中,t_1=[1986-1,1986-2],A={(a_1="华西附属二院")},该监测点可解释为:"监测于 1986-1~1986-2 期间,在华西附二院的病例".进一步来说,若 $|U_1(s_1)|$=480, $|U_1(s_2)|$=490,...,则说明该监测点在时间间隔 t_1,t_2 的病例数为 480,490,...由于在不同时间间隔 t_i 监测点的监测数值并不保持一致,按照定义 1,U_1 为出生缺陷数据集上以时间为维度的不确定监测点.

2.2 不确定干预评价模型

借助于定义 1,干预策略被施加在给定不确定监测点上:根据属性组合构成干预策略,数据实例的响应度用来衡量干预策略的强度(intensity),并用干预前后数据指标的变化来评价干预策略的效果(impact).干预策略为一组原子谓词组成的逻辑表达式,其形式化描述如下:

定义 2. 原子干预谓词(atomic intervention predication)是一个不可再分的逻辑判断式,记为 $\alpha\{A==c\}$;干预谓词域(intervention strategy space)是一组干预谓词的集合,记为 F;干预策略(intervention strategy)是由使用逻辑操作符($\wedge,\vee,\neg,\oplus$)连接起来的一组干预谓词组成,记为 s.若一条数据记录 d 满足干预策略 s,记为 $s(d)=true$;干预策略空间(intervention measure space)为一组干预策略的集合,记为 W.

图 1 给出了干预策略与干预谓词的逻辑结构示例.在出生缺陷项目背景下,一个具体的干预策略可能会被描述如下:"父母双方至少一方不为汉族".使用定义 1 中的描述规则,该干预策略可描述为 $s_1=\{\alpha_1\{mother\_ethinic!="汉"\}\vee\alpha_2\{father\_ethnic!="汉"\}\}$.

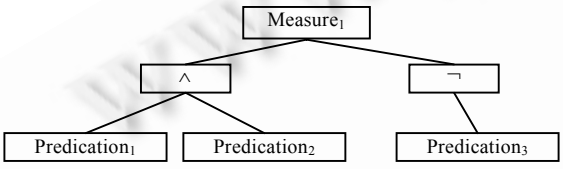


Fig.1 Architecture of intervention strategy and predications

图 1 干预策略、谓词结构

干预的强度由响应它的实例数来评定.考虑评定某策略干预强度,当政策制订后,若遵守该政策的人数比例大,则其强度大;否则其强度小.干预策略的效果由施加干预策略前后,观察数据在相应指标上的值的变化情况来确定.根据这个思路,干预强度及效果被定义如下:

定义 3. 设 U 为不确定监测点.给定时间间隔 t 和干预策略 s(s 的长度为其包含不同谓词数),t 内满足 s 与 s 内所有数据的实例数比值称为干预策略 s 在 t 的强度(intensity),记为 $\gamma_{s:t}$;给定评价算子 e,记数据项 u 满足 e 为 $e(u)=true$;称 t 内满足 s 所含评价算子与满足 s 的实例数比值为 s 在 U 的影响(impact),记为 $\theta_{s:t}$.

$$\gamma_{s:t}=\frac{\left|\{u\,|\,u_{time\_interval}=t,s(u)=true,u\in U\}\right|}{|U|} \tag{1}$$

$$\theta_{s:t}=\frac{\left|\{u\,|\,u_{time\_interval}=t,s(u)\times e(u)=true,u\in U\}\right|}{\left|\{u\,|\,u_{time\_interval}=t,s(u)=true,u\in U\}\right|} \tag{2}$$

实践中,通常外界因素使得相同的干预在不同时间段的效果有差异.为了精确地描述这种差异,我们提出了 p 概率干预策略,简称为 p-策略(p-strategy).

定义 4. 在不确定数据集 D 上,对不确定监测点 U:若干预策略 s_i 在给定的一系列时间段集合 $T=\{t_1,t_2,\ldots,t_n\}$ 中以概率 p 使其干预影响满足 $\theta_i\geqslant k$,则称 s_1 为限度为 k 的 p 概率干预策略,记为 p–$s_i.p$ 的计算由式(3)给出.

$$p_r(s)=\frac{\left|\left\{c\mid c\geqslant k, c\in\bigcup_{i=1}^{n}\theta_{s:t_i}\right\}\right|}{n} \tag{3}$$

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

例 2:设医学不确定数据集 D 上一个不确定监控点实例 U_1,其中 T={[1986-1,1986-2],[1986-3,1986-4],[1986-5,–1986-7],[1986-8,1986-9]},A={(a_1="华西附属二院")},对 U_1 施加干预策略 s_1,分别观察 T 中每个时间段上 s_1 的效果:假设有 3 次 s_1 的效果θ超过设定阈值 k,则 p=3/|T|=0.75.

观察 **1**. 设数据集 D 上一不确定监测点为 U={d| $\forall\, d.a_i\in A, d.time\_interval=t_i, \forall\, t_i\in T, d\in D$},评价 p 概率干预策略 p–s_1 时,随着|T|的增大,p–s_1 的评价开销 $COST(p–s_1)$增大.

证明:由评价 p–s_1 时,需要根据 S 中每个时间间隔对 U 进行一次评价,即

$$COST(p-s_1)=\sum_{t_i\in T}COST(s_1,U(t_i)) \tag{4}$$

可知,随着|T|的增大,$COST(p–s_1)$的取值增大.由定义 4,|T|越大,计算出 s_1 满足的概率阈值的准确度越高. □

由观察 1 可知,实际系统需要根据应用背景调整|T|以权衡系统的处理性能和精度.

2.3 问题定义

在前述概念基础上,我们实现了一个实验原型系统,能够在不确定性数据集上评价各种干预策略.主要解决了两个问题,即 I-to-E(intervention-to-evaluation)问题和 S-to-P(strategy-to-probability)问题.前者预测干预措施的可能的效果,后者计算给定干预策略达到确定效果的概率 p.下面给出形式化描述.

Problem-1(**I-to-E**):设干预策略空间 W 中包含 n 条干预策略,原始数据集为 D.任意给定空间中一条干预策略 s,从原始数据 D 中计算 s 的强度$\gamma_{s:t}$及影响$\theta_{s:t}$.

Problem-2(**S-to-P**):设干预策略空间 W 中包含 n 条干预策略,原始数据集为 D,任意给定空间中一条干预策略 s,在给定不确定监测点 U 计算 s 的影响$\theta\geqslant k$(k 为给定阈值)的概率 p.

根据前面的描述,干预策略评价问题可以分为两个步骤:(1) 计算数据是否满足给定干预策略;(2) 根据满足干预策略的数据计算评价算子值,获取干预影响及干预强度.

性能问题.当数据量很大时,系统的性能瓶颈主要会出现在这两种情况下:(1) 出现重复干预措施查询未能很好地复用之前的计算结果,造成反复计算浪费系统开销;(2) 多用户使用时重复计算公共谓词.

根据定义,干预策略由一组相关的干预原子谓词按照逻辑操作组合而成.给定一组干预策略,其谓词集合为 P={$p_1,p_2,p_3,\ldots,p_n$},评价其影响和强度在于评价流经系统的数据是否满足策略谓词集合中的谓词.主要瓶颈在干预谓词的评价计算,实验结果表明,不好的算法重复计算谓词评价占用了全部评价计算时间的 60%以上.本文设计数据结构存储干预空间中可能出现干预策略统计结果,再在此数据结构上评价干预策略,提高了效率.

影响谓词选择顺序的因素包括:(1) 选择率(selectivity):数据记录项匹配某给定谓词的概率;(2) 流行度(popularity):评价计划中包含该谓词的干预策略数量;(3) 开销(cost):一个项上计算某谓词所需的时间.

本文研究同一干预策略空间 W 中的干预策略共享公共谓词的评价结果,以提高系统的处理性能.在评价干预策略时,单条记录评价开销可忽略,本文的算法主要围绕谓词的选择率及流行度来优化谓词选择顺序.

3 干预策略评价优化算法

在给出干预策略评价的算法之前,首先分析本文提出干预策略的评价过程:

设 W 是给定的干预策略空间;S 是 W 内干预策略集合,即一组由所有干预策略中包含的干预谓词集合 $P(S)$,用于处理评价数据记录.策略 s_i 所包含的谓词集合记为 $Predicts$={$P(s_i)$|($\forall\, s_i\in S$)}.当对单条数据评价时,属于 $Predicts$ 集合的谓词必然有一个 *true* 或 *false* 的结果.又由于干预策略 s_1 是由 $Predicts$ 集合中的谓词组成,因此对单条数据评价时,干预策略 s_1 最终可化为一个有限命题.因为有限命题始终是可判定的,所以对单条数据,使用 s_1 评价总能给出判定结果.因此,使用 s_1 对不确定监测点 U 中数据实例进行评价即为统计 $\forall\, u\in U$ 在 s_1 上判定结

果的过程.评价干预策略前需要为公共谓词排序,即据数据情况,对 $P(s_1)$中的谓词排序,以提高系统的运行效率.图 2 展示了干预空间 W 的逻辑结构.

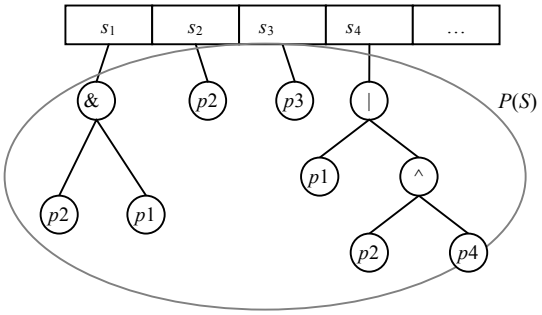


Fig.2 Architecture of an intervention strategy space

图 2 一个干预策略空间结构

3.1 *P*概率干预策略评价朴素算法

干预策略评价最直接的方法是按照策略的谓词对系统中每条数据过滤,使所有干预策略均为由谓词组成的过滤器.算法 1 给出了详细的描述.

算法 1. Naïve-evaluation.

输入:数据集 D,不确定观察点 U,时间间隔集 T,干预策略 s_i,阈值 k.

输出:干预强度$\gamma_{s:t}$,干预影响$\theta_{s:t}$,p.

步骤:

(1) p←0;
(2) **Foreach** (t **in** T) **do**
(3) { *defectNUM*←0; *totalNUM*←0; *coresNUM*←0;
(4) $\theta_{s:t},\gamma_{s:t}$←∅; //用于存储满足谓词的数据记录数
(5) U=*loadData*(D,t); //装载检测点数据
(6) **Foreach** (item u **in** U)**do** //对每个数据项 u
(7) { *totalNUM*++; //记录处理
(8) **if** (*impact_evaluate*(u,s)) **then** *defectNUM*++ //干预策略评价,满足返回 *true*,
(9) **if** (*intensity_evaluate*(u,s)) **then** *coresNUM*++;
(10) }
(11) $\theta_{s:t}$←*defectNUM*/*totalNUM*; $\gamma_{s:t}$←*coresNUM*/*totalNUM*; //输出在 $U(t)$评价结果
(12) **if** ($\theta_{s:t}\geqslant k$) **then** p++;
(13) }
(14) p←p/|T|;
(15) *Output*(p); //输出概率评价结果

Naïve-evaluate算法采用串型方式来评价单条干预策略.其优势是一遍扫描,评价单条干预策略时,效率较高.其缺点是,整个干预策略评价框架实现为一套系统,需要客户端提交海量的干预策略.每次执行算法 1 中第(6)、(7)步,需要重新对数据扫描以装载 $U(t)$的数据实例,需要较多的计算资源.实验结果表明,Naïve-evaluate 算法在干预策略评价时响应速度较慢,通常用户不能接受.为了解决这一问题,我们提出一个新的数据结构和相应的优化评价算法.

3.2 谓词统计树(predication statistical tree,简称PST)

由定义可知,干预策略由原子干预谓词的组合构成.实践表明,干预策略的统计中,通常有 30%~40%的相同

原子谓词.Naïve-evaluate 算法评价时,算法会重新计算每个干预策略中每个谓词的匹配.本节将减少重复计算.以提高算法处理效率.设干预策略集合为 $P(S)$.对其中所有谓词评价分为:(1) 处理原子谓词:根据谓词的综合代价对谓词排序,当数据记录到达时按该顺序评价;(2) 递归处理高阶谓词:根据计算量大小顺序及相关度评价剩余谓词.为了实现该计算目标,我们引入了谓词统计树(predication statistical tree,简称 PST).

定义 5(谓词统计树(predication statistical tree,简称 PST). 谓词统计树是一棵多叉树,由一个根节点(root node)以及层次节点(hierarchical node)的集合组成.PST 中所有节点的基本结构为 *Node*={*Label*,*Dimension*,*E*, *C*},其中 *Dimension*(维度集)是该层次节点反映的维度,*Label*(标签集)为该节点所代表的具体数值,*E*(元素集)是数据集中所有满足该节点的数据项的 ID,*C*(孩子集)为该节点的孩子节点集合.

例 3:图 3 展示了一个 PST 实例.根节点的元素集为空,一层节点的维度均为 *A*,二层节点的维度为 *B*,PST 的层次结构包含了不同维度、取值的统计信息.建立 PST 时,先按照维度取值生成第 1 层的节点,再根据下一个维度的取值生成下层节点.将相同维度及维度上取值的数据项纳入同一节点(向对应节点放入对应数据项 ID).

PST 建立后,就可以直接在 PST 上操作,以高效获取干预策略的评价结果.在图 3 中,若评价干预策略 $s_1=\{\alpha_1\{A=4\}\&\alpha_2\{B=2\}\}$,则在 PST 遍历至第 2 层节点时,若得到 *E*:{1,3},则可知满足该策略的数据项有 2 条.

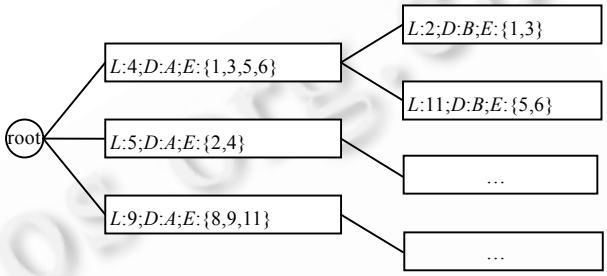


Fig.3 An instance of PST

图 3 一棵 PST 树的实例

PST 解决公共谓词重复评价的问题:在建立 PST 时,所用维度集由干预空间中所有原子谓词统计得到.谓词出现频率决定维度频率,根据维度频率和各维度中包含的不同数值量对维度排序,以确定 PST 节点的层次结构.同时,使低层次的节点为评价过程中访问最频繁的节点,缩短访问路径,以提高干预策略的评价速度.

对维度排序. 我们将维度的排序权值设置为 $unitprice(d)=p(d)\times n(d)$,其中,$p(d)$为维度的流行度,代表评价计划中包含谓词涉及该维度干预策略的比率;而 $n(d)$为该维度中不同数值的数量.在建立 PST 时,根据 *unitprice*(*d*),首先将维度集合从大到小排序.

3.3 PST创建算法及干预策略评价优化算法

干预策略评价优化算法分为两步:(1) 建立 PST 结构;(2) 干预策略评价.算法 2 给出了 PST 创建过程.

PST 算法首先根据干预策略集合统计维度的权值(即维度出现频率),再根据该权值对维度排序后放入维度集合.然后根据当前数据集中数据项增量生成 PST 树.

算法 2. PST-generate.

输入:数据流 D,$S=\{s_1,s_2,s_3,\ldots,s_k\}$,$P(S)=P(s_1)\cup P(s_2)\cup\ldots\cup P(s_k)$.

输出:完整 PST 树结构(从根节点 root 开始).

步骤:

(1) *Dimension*(*S*)=*split*(*S*); //分解所有干预策略,得到维度集合

(2) *Dimension*(*S*).*SortByPrice*( ); //按照维度权值 unitprice 排序

(3) *RecursiveCreateTree*(*root*,*Dimension*(*S*),*size*,*level*,*D*);} //递归构建 PST 树

(4) **Function *RecursiveCreateTree*(*root*,*dimension*,*size*,*level*,*DataSet*)**

(5) { **if** (*level*<0) **then return**;

(6) **else**{ //获取 *d* 维上不同值的数量,用以确定 *child* 数目

(7) *list*←*GetNominallist*(*DataSet*,*dimension*[*level*]);

(8) *list.Sort*( ); //按顺序排序

```
(9)          //为每个孩子节点创建
(10)         foreach (item in list) do {  //为当前节点加入数据,每个 node 对应的 list[n]值
(11)             newnode←new Node();
(12)             newnode.Value←item.value(); newnode.Dimension←dimension[level];
(13)           //创建新的数据集 newData,将 d 维等于 list[n]的数据项 id 记录进 newnode
(14)           newData←∅;
(15)           Foreach (dataitem in newData) do {
(16)               value←GetValue(dataitem,dimension[level]);  //获取对应维度取值
(17)               if (value=item.value) then {  //若 dataitem 在该维度取值满足节点值则添加
(18)                   newnode.Element.Add(GetID(dataitem));  //为节点添加 ID
(19)                   newData.Add(dataitem);  //为新数据集添加数据}      }
(20)           root.AddChild(newnode);  //继续为每个孩子节点创建树
(21)           RecursiveCreateTree(root.GetChild(),dimension,size,level-1,newData);}
(22)     }}
```

定理 1. 设总共有 N 个维度,即递归深度为 N,设每层的平均实例数为 P,数据集 $Dataset$ 的总规模为 M,则 PST-generate 算法的时间复杂度为 $O(P\times\mathrm{Log}(P)\times N)+O(M)$,空间复杂度为 $O(\mathrm{Log}_p(M))+O(M)=O(M)$.

证明:算法 2 中最大的系统消耗是第(6)、(7)两步,需要在当前数据子集中扫描指定维度中的不同实例数,然后对其排序(这里使用快速排序),因此,每层平均的计算次数为 $O(P\times\mathrm{Log}(P))$.由于对每个数据进行一遍扫描,计算次数为 $O(M)$,因此,总时间复杂度为 $O(P\times\mathrm{Log}(P)\times N)+O(M)$.该算法的空间消耗主要在两部分:每步递归产生的新数据子集,存储 PST 树型结构;由于算法中数据子集的产生被设计为从总数据量分流为小分支,因此,这部分的空间复杂度仍为 $O(M)$,而 PST 树型结构仅用节点保存了数据项的 ID,则这部分空间复杂度为 $O(\mathrm{Log}_p(M))$,总空间复杂度为 $O(\mathrm{Log}_p(M))+O(M)$,化简后,即 $O(M)$.证毕. □

例 4:设某段时间 t 内的数据集 $Dataset=\{d_1,d_2,d_3,d_4\}$,干预策略集合 $S=\{A,B,A\wedge B,A\wedge B,A\oplus B\}$.使用 PST-generate 建立 PST 树的过程是:

(1) 计所有原子谓词并根据其结果对维度集排序:$Dimension(S)=(B,A)$.

(2) 按维度 B 计算.设对应维度 B 的值分别是$\{d_1.B,d_2.B,d_3.B,d_4.B\}$,假设 $d_1.B=d_2.B=b_1,d_3.B=b_2,d_4=b_2$,则 $root$ 节点下分两个 $level2$ 子节点:$\langle B,b_1,\{d_1,d_1\}\rangle,\langle B,b_2,\{d_3,d_4\}\rangle$,此时,将 d_1,d_2 放入 $newDataset_1=\{d_1,d_2\}$;d_3,d_4 放入 $newDataset_2=\{d_3,d_4\}$.

(3) $newDataset_1,newDataset_2$ 按维度 A 计算:假设 $d_1.A=a_1,d_2.A=a_2;d_3.A=d_4.A=a_1$,则 $level3$ 子节点为$\langle A;a1;E:\{1\}\rangle,\langle A;a2;E:\{2\}\rangle,\langle A;a1;E:\{3,4\}\rangle$.

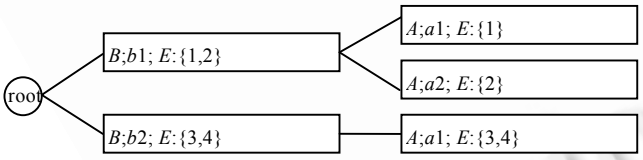


Fig.4 PST (PST_1) generated from Example 4

图 4 例 4 中生成的 PST 树 PST_1

(4) 根据本例的设置,PST 树最终的结构如图 4 所示.

在 PST 上评价干预策略.拥有了 PST 后,我们就获取到了一个关于数据集的统计信息.只需对 PST 遍历即可实现干预策略的评价.算法 3 给出了评价过程的细节.

算法 3. PST-evaluate.

输入:PST,不确定观察点 U 时间间隔集 T,干预策略 s_i,阈值 K.

输出:干预强度$\gamma s_i:t$,干预影响$\theta s_i:t,p$.

步骤:

(1) $p\leftarrow 0$; $nodeset\leftarrow\varnothing$

(2) **Foreach** (t in T) **do**

```
(3)     {   nodeset←trace(PST,dimension,s_i);   //找到对应节点集合
(4)         θs_i:t←getImpact(nodeset,e);   //输出干预影响(使用评价算子 e)
(5)         θs_i:t←getIntensity(nodeset);   //输出干预强度
(6)         if (θs_i:t>K) then p++;
(7)     }
(8)         p←p/|T|;
(1) Function trace(PST,s_i)
(2) {   nodeset←∅, filted_pathset←∅
(3)     P(s_i)←LoadPredication(s_i);   //获取 m_i 的谓词集合
(4)     pathset←DFSGetPathSet( );   //深度优先遍历 PST 获取每条路径
(5)     Foreach (p in pathset) do   //过滤 pathset
(6)     {If (evaluate(s_1,p)=true) Then   //若该条路径满足干预策略 s_1 的逻辑,则放入 filted_pathset
(7)         filted_pathset←filted_pathset∪p;}
(8)     nodeset←GetNode(filted_pathset)   //获取指定的节点
(9)     return nodeset;}
```

定理 2. 设总数据量为 N,PST 的每个节点中包含的 element 的平均数为 M,且设 PST 为 P-叉树(假设 PST 每个节点平均有 P 个孩子节点),则算法 PST-evaluate 的时间复杂度为 $O(N/M)$,空间复杂度为 $O((N/M)/\mathrm{Log}_P(N/M))$(该算法空间复杂度小于 $O(N/M)$).

证明:由题设,PST 所有的节点数为 N/M,则深度遍历 PST 的时间复杂度为 $O(N/M)$.过滤从 PST 中获取到 *pathset* 中路径的时间复杂度亦为 $O(N/M)$.由于 *nodeset* 很小,其对强度的影响可忽略.算法 PST-evaluate 的时间复杂度为 $O(N/M)$.由题设可知,PST 平均高度为 $\mathrm{Log}_P(N/M)$.算法运行过程中,空间消耗最大的是存储深度优先遍历后得到的 *pathset*,其空间复杂度为 $O((N/M)/\mathrm{Log}_p(N/M))$,因此,算法总空间复杂度为 $O((N/M)/\mathrm{Log}_p(N/M))$. □

直观上,上述算法空间复杂度比 $O(N/M)$小常数倍.如,当 P=2,N/M=1024 时,复杂度约估计为 O(1024/10).

例 5:设生成的 PST_1 树如图 4 所示.采用 PST-evaluate 评价干预策略集合 $S=\{B(B=b_1),!B(B=b_2),A|B(B=b_1,a=a_1),A\wedge B(B=b_2,a=a_1)\}$.在本例中,排序维度集合为 $dimension=\{B,A\}$.假设 a_1 表示缺陷,a_2 表示无缺陷.

评价 $s=!B(B=b_2)$.选取所有 $B\neq b_2$ 的节点,即节点$\langle B,b_1,E\{1,2\}\rangle$.因此,干预强度为$\gamma s:t$=2/4=0.5,干预影响因子集合为 $s:t$=0.25/0.5=0.5.实验结果表明,如果使用过多的与操作,则会造成满足样本数较小,使统计结果失效.在实际策略构建时应当避免使用过于严格的谓词组合,由此,通常长度上限为 10 的策略比较适合.

4 实验及性能评价

4.1 实验配置及数据集特征

数据集特征.来自中国出生缺陷监控中心 1986 年~1991 年的生育数据,数据均来自磁盘分段文件,不同文件中数据模式不同.为了解决数据模式非结构化问题,我们将数据转化为“infant;$\mathrm{attribute}_1$,value_1;$\mathrm{attribute}_2$,value_2;$\mathrm{attribute}_3$,value_3…”模式,其中,infant 是婴儿的记录编号;attribute 和 value 分别代表该条记录的属性编号与记录在该属性上的值.例如,“70397;129,2;130,1;131,4;132,1;133,2;134,1;205,1;209,2;211,3050;212,39;213,2;214,3;215,1;216,2;217,2;218,1;219,1;220,2;221,3;224,11;238,434;239,1986;240,10;241,1;243,1;244,370000;245,370100”.处理后,共有 969 476 条数据记录.表 2 列出了本部分使用数据集的相关信息.

Table 2 A summary of experiment datasets

表 2 实验数据集描述

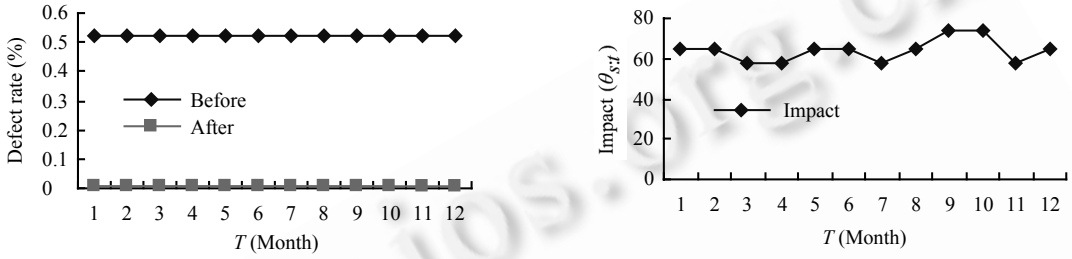
	Birth defect data set	Synthetic data set
Infants number	969 476	1 000 000
Maximum dimension number	245	500
Strategy number	4	4 000

实验平台.机器配置:CPU:Intel® Pentium® Dual E2160;@1.80GHZ;RAM:2GB,使用 Visual Studio 2008 的 C# 开发平台.评价原型系统包括随机干预策略产生器,评价对比平台,且分别实现了 Naïve-evaluate 算法、PST-generate 算法和 PST-evaluate 算法以进行比较实验.

4.2 正确性验证实验

首先使用干预策略 s={α_1{*mother_ethnic*!="汉"} ∨ α_2{*father_ethnic*!="汉"}}在不确定维度 T 上进行正确性验证,策略 s 对应"现实生活中医生希望查找父母双方有一方不是汉族时,小孩出生时的缺陷概率变化情况".由于 Naïve-evaluate 算法是直接对原始数据过滤处理,其计算的结果作为验证正确性的基础.

干预策略评价测试分别在两种情况下进行:(1) 干预策略评价;(2) 概率干预策略评价(包含概率计算);实验对比二者在处理数据不确定性后缺陷精度的变化及分别按照阈值 k 取值的不同来分析计算处策略 s 在不同数据集中的概率 p.正确性实验使用了 1988 年的出生缺陷数据(共 16 011 条数据),T 分别设置为 1988 年 1 个月~12 个月.获取干预策略施加前后缺陷率及影响效果如图 5 所示.



(a) Defect rate results before and after applying intervention strategy s
(a) 施加干预策略 s 前后获取的缺陷率

(b) Impact value (θ_s) of intervention strategy s
(b) 干预策略 s 的影响值θ_s

Fig.5 Result of algorithm Naïve-evaluate and PST-evaluate on real data in 1988
(Impact values of Naïve and PST algorithm are the same)

图 5 1988 年数据上,Naïve-evaluate, PST-evaluate 计算干预策略 s 结果(Naïve,PST 方法完全相同)

通过对 1988 年数据进行计算,Naïve-evaluate 方法和 PST-evaluate 方法得到的结果完全一致.由于 Naïve-evaluate 方法是直接对全部数据进行顺序扫描,获取的结果的精确度具有非常高的可信度,这表明,经过建立 PST 树,再从上进行干预策略评价的优化算法 PST-evaluate 的计算是完全正确的.

根据阈值 k 取值的不同,可以得到干预策略达到既定影响的不同概率.本实验的计算结果见表 3.

Table 3 Impact of different values of k to the intervention strategy probability of s

表 3 不同 k 取值对干预策略 s 概率的影响

k value	Probability p of s
60	0.33
70	0.83
75	1

表 3 表明,干预策略 s 在 k 取 70 时,其产生影响 70 以上的概率为 0.83.有了干预策略达到一定影响的概率值,不仅可以比较其影响值的大小,而且可以在给定 k 值的情况下,比较不同策略可能产生一定影响的概率.k 值通常可以根据医学专业人员的常识来指定,这使得算法评价的结果更加符合实际情况.

知识发现实例:由图 5 中的实验结果可以看出,干预策略 s={α_1{*mother_ethnic*!="汉"}∨α_2{*father_ethnic*!="汉"}}在施加前的出生缺陷率为α_b=0.52,根据处理结果,获取干预后出生缺陷率分别为:

$\alpha_{a\&month=1}$=0.00805, $\alpha_{a\&month=2}$=0.0079, $\alpha_{a\&month=3}$=0.0095, $\alpha_{a\&month=4}$=0.0089, $\alpha_{a\&month=5}$=0.0082,
$\alpha_{a\&month=6}$=0.0081, $\alpha_{a\&month=7}$=0.0092, $\alpha_{a\&month=8}$=0.0080, $\alpha_{a\&month=9}$=0.0074, $\alpha_{a\&month=10}$=0.0074,
$\alpha_{a\&month=11}$=0.0089, $\alpha_{a\&month=12}$=0.0076,

则缺陷率相对变化率τ为

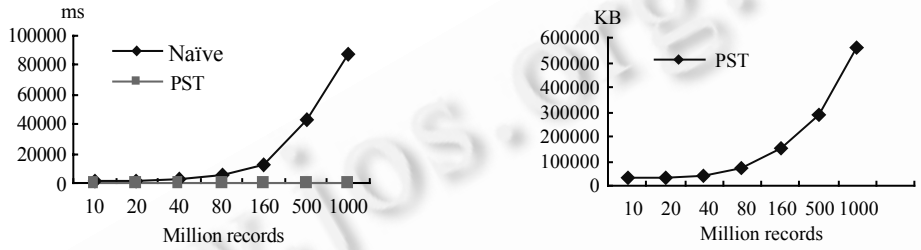
$$\gamma=\frac{\sum_{i=1}^{12}(\alpha_b-\alpha_{a\&month=i})}{\sum_{i=1}^{12}\alpha_{a\&month=i}}=62.93 \tag{5}$$

也就是说,若夫妇双方有一方不是汉族,则生育小孩的出生缺陷率低于正常情况的 60 倍以上.这一结果与四川大学华西附二院的医学观察的结果吻合[25].此外,我们对策略 $s_1=\{\alpha_1\{mother\_age\geqslant 20\}\vee\alpha_2\{mother\_age\leqslant 30\}\}$ 及策略 $s_2=\{\alpha_1\{mother_age\geqslant 30\}\}$进行了比较,比较结果为 $Impact(s_1)=0.174$,$Impact(s_2)=0.125$,说明施加干预策略 s_1 后获得的影响强于 s_2.因此说明,在对缺陷率影响主题下,20 岁~30 岁年龄段妇女生育的出生缺陷率较低,这一点符合医学文献[25,26]中的相关描述.这说明本文模型具有实际的医学意义.

4.3 性能测试

在第 4.2 节中已对干预分析模型以及两个评价算法进行验证.本节将分析两种算法的性能,为实践提供指导性意见.对真实数据集中实例进行复制,得到 1 000 000 条仿真数据,用于测试算法性能.

单干预策略评价.首先对单条干预策略评价过程中算法的性能进行测试.我们沿用第 4.2 节中的干预策略 s,Naïve-evaluate 及 PST-evaluate 算法的评价性能如图 6 所示.



(a) Running time efficiency comparison of Naïve and PST algorithm
(a) Naïve 和 PST 算法运行时间效率的比较

(b) Space efficiency of PST algorithm
(b) PST 算法空间效率

Fig.6 Comparison of running efficiency of the algorithm Naïve-evaluate and PST-evaluate
图 6 Naïve-evaluate 与 PST-evaluate 算法性能测试比较

图 6 中每个测试数据均是处理 10 次后的平均值:Naïve-evaluate 是顺序扫描,其时间复杂度随着数据量的增大而增加;而由于采用了 PST 树,评价效率提高,几乎不受数据量的影响,每次评价 3ms.这表明,PST 树极大地提高了 p-干预策略的挖掘效率.PST 随着数据量的增大而增加,最大为 558 173KB,可全放入内存,从而提高了评价效率.

海量干预策略评价.本实验随机创建了 4 000 条干预策略(随机产生长度限制为 10),让它们的评价计算同时在原型系统上运行.由于 Naïve-evaluate 算法的性能不足以处理多干预策略并发处理,本部分实验仅测试 PST-evaluate 算法在海量干预策略并发评价时的性能.

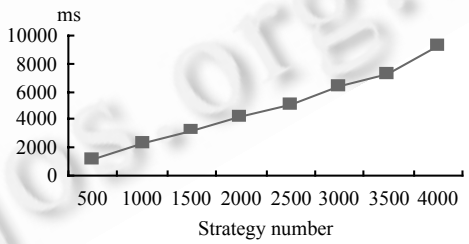


Fig.7 Relation between evaluating performance and number of intervention strategy
图 7 干预策略数和评价性能的关系

由于 PST-evaluate 算法的每次评价不受数据量的影响.实验结果表明,并发评价 4 000 个干预策略的时间效率在 8 131ms~11 324ms 之间.图 7 中列出了随着干预策略数的增加,系统的时间开销情况.可以看出,本文算法的性能随着干预策略的增加,基本保持线性变化,基本上满足了系统上线后并发查询的需求.

由于算法的空间开销主要来源于所构建的 PST 树,因此,PST 树的空间消耗不变.以上分析说明了 PST-evaluate 在处理海量干预策略时的优势,达到了算法设计的目标.在不确定干预评价系统中,系统采用离线计算建立 PST 树,并将其物化到硬盘,仅在统计到干预策略集发生大幅度变化(一段时间内用户在使用干预评价

系统时提交的干预策略)时才重新生成 PST 树,进一步地提高了系统运行效率.

5 结束语

本文总结了干预分析相关的工作,在 what-if 分析以及不确定数据分析的基础上,提出不确定干预策略的分析模型,并给出挖掘算法及优化评价算法.文章还重点解决了在百万级数据量上评价并发评价海量干预策略的问题,为实现实际评价系统提供了指导性的实验结果.然而,关于在不确定数据上进行干预知识的发现还刚刚起步,很多工作需要深入展开.未来的工作方向是,高效搜索优化干预策略,给定概率 p 产生搜索满足的干预策略.即指定概率 p_0 和影响 θ,计算所有能够得到以概率 p_0 达到改变 θ 效果的干预策略集合,作为决策备选方案.

References:

[1] Goebel M, Gruenwald L. A survey of data mining and knowledge discovery software tools. ACM SIGMOD Record, 1987,16(3): 34−48. [doi: 10.1145/38714.38724]

[2] Tang CJ, Zhang Y, Tang L, Li C, Chen Y. Survey on mining kinetic intervention rule from sub-complex systems. Computer Applications, 2008,28(11):2732−2736, 2748 (in Chinese with English abstract).

[3] Zhou AY, Jin CQ, Wang GR, Li JZ. A survey on the management of uncertain data. Chinese Journal of Computers, 2009,32(1): 1−16 (in Chinese with English abstract).

[4] Nierman A, Jagadish HV. ProTDB: Probabilistic data in XML. In: Bernstein PA, Ioannidis YE, Ramakrishnan R, Papadias D, eds. Proc. of the 28th Int'l Conf. on Very Large Data Bases. Morgan Kaufmann Publishers, 2002. 646−657.

[5] Cavallo R, Pittarelli M. The theory of probabilistic databases. In: Stocker PM, Kent W, Hammersley P, eds. Proc. of the 13th Int'l Conf. on Very Large Data Bases. Morgan Kaufmann Publishers, 1987. 71−81.

[6] Fuhr N, Rölleke T. A probabilistic relational algebra for the integration of information retrieval and database systems. ACM Trans. on Information Systems, 1997,15(1):32−66. [doi: 10.1145/239041.239045]

[7] Chau M, Cheng R, Kao B, Ng J. Uncertain data mining: An example in clustering location data. In: Ng WK, Kitsuregawa M, Li JZ, Chang KY, eds. Proc. of the 10th Pacific-Asia Conf. on Knowledge Discovery and Data Mining. Heidelberg: Springer-Verlag, 2006. 199−204. [doi: 10.1007/11731139_24]

[8] Green TJ, Tannen V. Models for incomplete and probabilistic information. IEEE Date Engineering Bulletin, 2006,29(1):17−24. [doi: 10.1007/11896548_24]

[9] Abiteboul S, Senellart P. Querying and updating probabilistic information in XML. In: Bertino E, Christodoulakis S, Plexousakis D, Christophides V, Koubarakis M, Böhm K, Ferrari E, eds. Proc. of the 9th Int'l Conf. on Extending Database Technology: Advances in Database Technology. Heidelberg: Springer-Verlag, 2006. 1059−1068. [doi: 10.1007/11687238_62]

[10] Senellart P, Abiteboul S. On the complexity of managing probabilistic XML data. In: Libkin L, ed. Proc. of the 26th ACM SIGMOD-SIGACT- SIGART Symp. on Principles of Database Systems. New York: ACM, 2007. 283−292. [doi: 10.1145/1265530.1265570]

[11] Tao YF, Cheng R, Xiao XK, Ngai WK, Kao B, Prabhakar S. Indexing multi-dimensional uncertain data with arbitrary probability density functions. In: Böhm K, Jensen CS, Haas LM, Kersten ML, Larson PÅ, Ooi BC, eds. Proc. of the 31st Int'l Conf. on Very Large Data Bases. New York: ACM, 2005. 922−933.

[12] Pei J, Jiang B, Lin XM, Yuan YD. Probabilistic skylines on uncertain data. In: Koch C, Gehrke J, Garofalakis MN, Srivastava D, Aberer K, Deshpande A, Florescu D, Chan CY, Ganti V, Kanne CC, Klas W, Neuhold EJ, eds. Proc. of the 33rd Int'l Conf. on Very Large Data Bases. New York: ACM, 2007. 15−26.

[13] Box GEP, Jenkins GM, Reinsel GC. Time Series Analysis, Forecasting and Control. 3rd ed., Englewood Clifs: Prentice Hall, 1994.

[14] Box GEP, McGregor JF. The analysis of closed-loop dynamic stochastic systems. Technometrics, 1974,16(3):391−398.

[15] Box GEP, Tiao GC. Intervention analysis with applications to economic and environmental problems. Journal of the American Statistical Association, 1975,70(349):70−79. [doi: 10.2307/2285379]

[16] Xiong Y, Yu S. Application of intervention analysis by ARMAX. Statistics and Decision, 2003,(11):17−26 (in Chinese with English abstract).

[17] Zhu YY, Shasha D. Efficient elastic burst detection in data streams. In: Getoor L, Senator TE, Domingos P, Faloutsos C, eds. Proc. of the SIGKDD 2003. NewYork: ACM, 2003. [doi: 10.1145/956750. 956789]

[18] Ihler A, Hutchins J, Smyth P. Adaptive event detection with time—Varying Poisson processes. In: Eliassi-Rad T, Ungar LH, Craven M, Gunopulos D, eds. Proc. of the SIGKDD 2006. New York: ACM, 2006. 207−216. [doi: 10.1145/1150402.1150428]

[19] Cai YD, Clustter D, Pape G, Han JW, Welge M, Auvil L. MAIDS: Mining alarming incidents from data streams. In: Weikum G, König AC, Deßloch S, eds. Proc. of the ACM SIGMOD. New York: ACM, 2004. 919−920. [doi: 10.1145/1007568.1007695]

[20] Aggarwal CC, Yu PS. Outlier detection with uncertain data. In: Apte C, *et al*., eds. Proc. of the SIAM Data Mining Conf. 2008. 483−493.

[21] Zhou XY, Sun ZH, Zhang BL, Yang YD. A fast outlier detection algorithm for high dimensional categorical data streams. Journal of Software, 2007,18(4):933−942 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/18/933.htm [doi: 10.1360/jos180933]

[22] Wang Y, Tang CJ, Li C, Chen Y, Yang N, Tang R, Zhu J. Intervention events detection and prediction in data streams. In: Li Q, Feng L, Pei J, Wang XYS, Zhou XF, Zhu QM, eds. Proc. of the Joint Int'l Conf. on Asia-Pacific Web Conf. (APWeb) and Web-Age Information Management (WAIM). Heidelberg: Springer-Verlag, 2009. [doi: 10.1007/ 978-3-642-00672-2_45]

[23] Lakshmanan LVS, Russakovsky A, Sashikanth V. What-If OLAP queries with changing dimensions. In: Alonso G, Blakeley J, eds. Proc. of IEEE the 24th Int'l Conf. on Data Engineering. Washington: IEEE, 2008. 1334−1336. [doi: 10.1109/ICDE.2008.4497547]

[24] Wang S, Xiao YQ, Zhang YS, Chen H. Research on OLAP system supporting what-if analysis. Chinese Journal of Computers, 2008,31(9):1573−1586 (in Chinese with English abstract).

[25] Dai L, Zhu J. Dynamic monitoring of neural tube defects in China during 1996 to 2000. Chinese Journal of Preventive Medicine, 2002,36(6):402−405. (in Chinese with English abstract).

[26] Liu XY, Zeng WQ, Liu JP. Analysis of factors associated with defects of birth defects. Journal of Chinese Current Clinical Medicine,2004,2(8):458−459 (in Chinese with English abstract).

附中文参考文献:

[2] 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.

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

[16] 熊焰,余石.干预分析的 ARMAX 模型及应用.统计与决策,2003,(11):17−26.

[21] 周晓云,孙志挥,张柏礼,杨宜东.高维类别属性数据流离群点快速检测算法.软件学报,2007,18(4):933−942. http://www.jos.org.cn/1000-9825/18/933.htm [doi: 10.1360/jos180933]

[24] 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.

[25] 代礼,朱军.1996~2000 年全国神经管缺陷的动态监测.中华预防医学杂志,2002,36(6):402−405.

[26] 刘小毓,曾维钦,刘建平.围产儿出生缺陷相关因素分析.中华现代临床医学杂志,2004,2(8):458−459.

王悦(1981－),男,四川成都人,博士,主要研究领域为数据库,知识工程.

李红军(1977－),男,博士生,主要研究领域为数据挖掘.

唐常杰(1946－),男,教授,博士生导师,CCF 高级会员,主要研究领域为数据库与知识工程,数据 挖掘.

郑皎凌(1981－),女,博士,主要研究领域为数据挖掘.

杨宁(1974－),男,博士,主要研究领域为数据挖掘.

朱军(1964－),女,博士,教授,主要研究领域为出生缺陷干预.

张悦(1983－),女,硕士,主要研究领域为数据挖掘.