文章编号:1001 - 9081(2007)10 - 2481 - 03

基于时事折换图的因果知识量化分析方法

郑 伟,孟晓风,孙 群
(北京航空航天大学 仪器科学与光电工程学院,北京 100083)
(zw3475@163.com)

摘 要:针对目前协同因果事件族的知识分析缺乏时态信息和量化手段等问题,提出基于时事折换图的因果知识量化分析方法。首先运用时事折换思想建立包含时态信息的因果知识事件域描述机制,然后考虑原因事件对结果事件发生冲击影响和时间累积影响,提出因果度、依赖度、影响度三种指标并推出其计算公式,实现了对因果知识的量化描述和双向推理。应用实例表明该方法具有高效、细致、直观优点。

关键词:时事折换图;因果知识;量化分析;因果度
中图分类号: TP18 **文献标志码**:A

Quantitative analysis method for causality knowledge based on time-event exchange diagram

ZHENG Wei, MENG Xiao-feng, SUN Qun
(*School of Instrument Science and Optoelectronics Engineering, Beijing University of Aeronautics and Astronautics, Beijing* 100083, *China*)

Abstract: For the void of time information and quantitative methods of the knowledge analysis for conjunction causality events, a quantitative analysis method based on time-event exchange diagram was proposed. Firstly the time-event exchange thought was adopted to establish the description mechanism that contained time information and based on event domain of causality knowledge. Then the happening and cumulate time influences of the cause event to the result event are considered. Afterwards the causality-degree, rely-degree and effect-degree including their calculating formulas are introduced which realize the quantitative description and bidirectional reasoning of causality knowledge. An example of application demonstrates that the method has high efficiency, elaborate and visual advantages.

Key words: time-event exchange diagram; causality knowledge; quantitative analysis; causality-degree

0 引言

协同因果事件族是广泛存在于人类生活的一类事件信息系统,在信号处理、故障诊断,案情剖析、病理推断等诸多领域都有其存在形态,其特点表现为多个原因事件在非一致时域区间的协同作用促使结果事件的发生或结束[1]。

目前对因果事件信息系统表达及推理手段主要基于概率论,如贝叶斯网、因果图等。贝叶斯网是一种有向无环的图形结构,可用来表述事物之间概率因果关系,广泛用于故障诊断、病理推断等领域[2,3]。因果图是在传统贝叶斯网基础上引入布尔逻辑运算发展出的不确定性知识推理模型,克服了贝叶斯网不能处理因果环路结构、没有考虑条件概率随时间动态变化等不足处[4]。以上述方法为基础,延拓出一些多技术融合方法,例如利用贝叶斯网结合遗传算法分析医学数据[5],在因果图中引入模糊推理算法扩展其模糊量处理能力[6]等。

上述方法通常以条件概率为建模基础,模型中缺乏因果事件的时态信息。文献[7]利用数据挖掘技术实现了对因果事件时间序列关联模式的发现,但仅完成双时间序列的事件变动频繁模式发现。文献[8]将时态信息融入因果关系决策树分析,但由于采用绝对时间记录并表述因果事件,处理多序列高精度时间采样将面临较大实现压力。

为增强决策系统获取因果知识的全面性、细致性,本文首先展开时间/事件的内联关系研究,采用时事折换思想,提出原因/结果事件的因果度、依赖度、影响度指标,对协同因果事件族的因果知识进行了量化描述和双向推理,最后以应用实例说明了该方法的使用效果。

1 协同因果事件族数学描述

首先对因果事件信息系统的结构刻画如下:

定义 1 一个因果事件信息系统可表示为:

$S = \langle T, C, R, V, f \rangle$

$T = \{t_1, \cdots, t_k\}$ 是时域集合; $C \cup R = D$ 是事件集合,子集 $C = \{C_1, \cdots, C_m\}$ 和 $R = \{R_1, \cdots, R_n\}$ 分别称为原因事件集和结果事件集; $V = \cup v_d$, $d \in D$ 是事件状态集合, v_d 表示某个事件 $d \in D$ 的状态, $V = \{$发生态,结束态,延续态,消失态$\}$; f 定义一个信息函数, $f: T \times D \rightarrow V$,它指定 T 中每一时域元素 t 所在位置的事件状态。

定义 2 当用 T 表述 D 中元素状态时, $T \mid D$ 表示根据 D, T 中的时域元素构成的描述 D 的时域区间族,记为 $T \mid D =$

收稿日期:2007 - 05 - 21;**修回日期**:2007 - 07 - 7。

作者简介:郑伟(1975 -),男,重庆人,博士研究生,主要研究方向:嵌入系统测控、人工智能、数据挖掘技术; 孟晓风(1955 -),男,重庆人,教授,博士生导师,主要研究方向:检测与诊断技术、测控系统及软件开发平台; 孙群(1979 -),男,山东人,博士研究生,主要研究方向:计算机测控系统。

$[TB, TE]$,其中,TB 是 D 发生时刻族,TE 是 D 结束时刻族。

对 $R_j \in R(j = 1, 2, \cdots, n)$, $C_i \in C(i = 1, 2, \cdots, m)$,若 C_i 和 R_j 存在因果关系,记为 $C_i \mapsto R_j$。C_i 和 R_j 满足如下定理。

定理 1 时域区间因果关系判定定理。

$\exists T \mid C_i = [t_p, t_q], T \mid R_j = [t_g, t_h] (i = 1, 2, \cdots, m; j = 1, 2, \cdots, n; t_p, t_q, t_g, t_h \in T)$, $C_i \mapsto R_j$, 必要非充分条件是 $[t_p, t_q] \cap [t_g, t_h] = \varnothing$。

证明:

1) 必要性证明

若 $C_i \mapsto R_j$,则 $f([t_p, t_q], C_i) \rightarrow v_{R_j} = =$ {发生态} ‖ {结束态};

当 $f([t_p, t_q], C_i) \rightarrow v_{R_j} = =$ {发生态},有: $t_p \leqslant t_g \leqslant t_q$;

当 $f([t_p, t_q], C_i) \rightarrow v_{R_j} = =$ {结束态},有: $t_p \leqslant t_h \leqslant t_q$,

上述两种情况下均有: $[t_p, t_q] \cap [t_g, t_h] \neq \varnothing$。

$\therefore C_i \mapsto R_j \Rightarrow [t_p, t_q] \cap [t_g, t_h] \neq \varnothing$

2) 非充分性证明

设 $T \mid C_i^* = [t_p^*, t_q^*], T \mid R_j = [t_g, t_h], T \mid C_i = [t_p, t_q]$, 由必要性证明结论知:

$C_i^* \mapsto R_j \Rightarrow [t_p^*, t_q^*] \cap [t_g, t_h] \neq \varnothing$。

若有 $[t_p^*, t_q^*] \subset [t_p, t_q]$,则有: $[t_p, t_q] \cap [t_g, t_h] \neq \varnothing$。

由前提假设 $C_i \not\mapsto R_j$ 知:

$[t_p, t_q] \cap [t_g, t_h] \neq \varnothing \not\Rightarrow C_i \mapsto R_j$。

综 1)、2) 所证, $C_i \mapsto R_j$ 必要非充分条件是:

$[t_p, t_q] \cap [t_g, t_h] \neq \varnothing$。

从定理 1 获取的直观认识是:存在因果关系的事件可以在时域中寻求其支持知识,但反之,时域中的支持知识并不足以说明事件间必然存在因果关系。这种认识启发我们可以建立时序和事件的映射关系,通过时序的关联信息实现对事件因果关系的解析并进一步结合其他知识进行决策推理。

在定理 1 给定条件下,有如下推论成立:

推论 1 $[t_p, t_q] \cap [t_g, t_h] \neq \varnothing \Rightarrow C_i \not\mapsto R_j$。

证明 由定理 1 必要性证明知, $C_i \mapsto R_j \Rightarrow \lceil [t_p, t_q] \cap [t_g, t_h] \neq \varnothing$,其逆否命题亦即推论 1 也应成立。

推论 1 的意义是可以利用该结论实现对协同因果事件族的关系约简,删除不存在因果关系的事件组合,尤其适用于高维原因 / 结果事件序列的关系化简。

2 基于时事折换图的因果知识量化分析

2.1 时事折换图

本文提出的时事折换图基本思想是将时间域度量折换为事件域度量,实现从连续量向离散量的转化,以便进行事件相关性的数据挖掘。由定义 1、2 易知,定义 1 所述的信息函数 f 实际是一种映射函数,实现从时序族向事件族的如下映射:

$f(TB, D) \rightarrow$ {发生态} $\in V$,

$f(TE, D) \rightarrow$ {结束态} $\in V$,

$f((TB, TE), D) \rightarrow$ {延续态} $\in V$,

$f((-\infty, TB) \cup (TE, +\infty), D) \rightarrow$ {消失态} $\in V$。

基于上述内联关系建立折换图模型,步骤如下:

1) 知识学习。获取因果事件时域区间族 $T \mid D = [TB, TE]$,绘制时序图 G。

2) 初始设置。设时间指针为 $timer$,事件频度指针为 x_k,初始化 $timer = 0$, $x_k = 0$。

3) 时域扫描。以 $timer = timer + \Delta t (\Delta t \rightarrow 0)$ 扫描 G 时间轴 $(0, +\infty)$。

4) 特征捕捉。若 $f(timer, D) \rightarrow v_d = =$ {{发生态} ‖ {结束态}}, $d \in D$,则 $x_k = x_k + 1$, 标定事件 d 频度坐标为 x_k。5) 返回执行 3) 直至 $timer = = G$ 结束时刻。

经过上述步骤,原事件的时间区间度量指标折换为和该事件相关的事件发生 / 结束的频度指标。以一个 8 输入,3 输出的信号系统为例,当把输入信号看作原因事件,输出信号看作结果事件时,时事折换图见图 1。

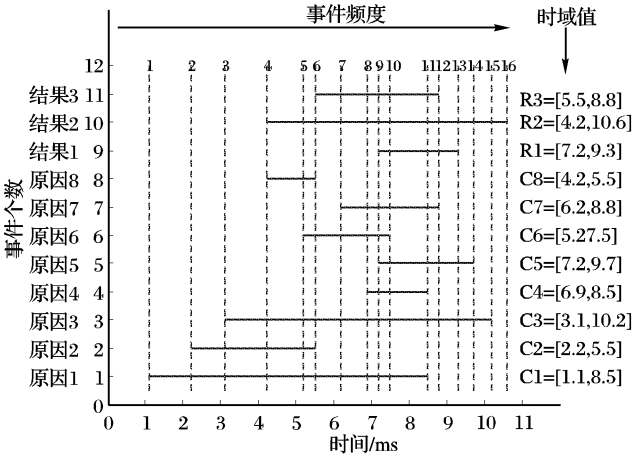

图 1 时事折换图图例

该方法有以下特点:

1) 不依赖时间采样精度,即使无法得到明确的时域信息,根据事件集合发生 / 结束的相对关系也可描绘出事件频度图。

2) 事件频度融合了时间信息和关系信息,表述方式比时域表述法更为简单,是一种整数度量方法。

3) 从数据的时态结构中挖掘原因 / 结果事件的关联性,知识发现阶段不需先验知识。

2.2 量化分析

时事折换后对时域对象的描述变为: $\exists T \mid C_i = [x_p, x_q], T \mid R_j = [x_g, x_h] (i = 1, 2, \cdots, m; j = 1, 2, \cdots, n, 1 \leqslant p, q, g, h \leqslant k)$, C_i 和 R_j 之间存在以下几种典型关系,见图 2。

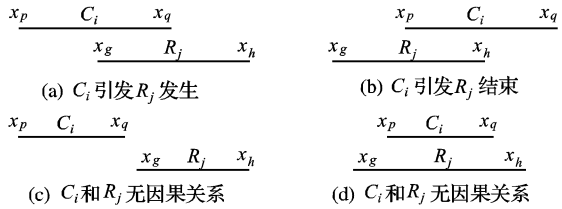

图 2 C_i 和 R_j 典型关系

考虑原因事件对结果事件的发生冲击影响和时间累积影响。针对图 2(a), C_i 发生点与 R_j 发生点之间其他事件出现的频度越低,表明 C_i 发生动作对 R_j 的影响越大,引入 C_i 对 R_j 的冲击影响函数:

$$k_1 / (x_g - x_p) \tag{1}$$

k_1 为沿影响系数, $0 < k_1 \leqslant 1$。

C_i 与 R_j 重叠区域出现其他事件的频度越高,表明 C_i 对 R_j 的时间累积影响越大,引入 C_i 对 R_j 的时间累积影响函数:

$$k_2 (x_q - x_g) / (x_h - x_g) \tag{2}$$

k_2 为区间影响系数, $0 < k_2 \leqslant 1$。

至此,给出事件因果度定义:

定义3　$\forall T|C, T|R$,以 $C_i(i=1,2,\cdots,m)$ 为行向量,$R_j(j=1,2,\cdots,n)$ 为列向量建立二维矩阵 $\sigma_{m\times n}$,$\sigma_{m\times n}$ 中第 i 行 j 列元素数值反映 C_i 和 R_j 因果关系的强弱,记为因果度 σ_{ij}。

按图2(a)推理思路,结合时频图以及对临界点的考虑,推出 σ_{ij} 计算公式:

$$\sigma_{ij}=\begin{cases}k_1\dfrac{1}{0.5}+k_2\dfrac{x_q-x_g}{x_h-x_g}, & x_p=x_g\\ k_1\dfrac{1}{x_g-x_p}+k_2\dfrac{x_q-x_g}{x_h-x_g}, & x_p<x_g\leqslant x_q\\ k_1\dfrac{1}{x_h-x_p}+k_2\dfrac{x_h-x_p}{x_h-x_g}, & x_g<x_p<x_h\leqslant x_q\\ k_1\dfrac{1}{0.5}, & x_p=x_h\\ 0, & \text{其他}\end{cases}\tag{3}$$

因果度指标实现了对因果事件族可能存在的因果关系的量化描述,但正如定理1提供的信息,这种量化作用是导向性和启发性的,主要起到知识发现的作用。为实现决策规则的知识推理,本文引入条件概率推出事件依赖度和影响度指标。

定义4　$\exists R_j\in R(j=1,2,\cdots,n)$,$R_j$ 发生或结束依赖于集合 C 中各元素协同作用,对 $C_i(i=1,2,\cdots,m)$ 而言,R_j 对 C_i 的依赖程度记为依赖度 $RY_i(R_j)$。

当用 R_j 对集合 C 中各元素的概率因果度平方和的开方值表示协同合力时,有:

$$RY_i(R_j)=\frac{P(C_i|R_j)\sigma_{ij}}{\sqrt{\sum_{i=1}^{m}[P(C_i|R_j)\sigma_{ij}]^2}}\tag{4}$$

其中,$\sum_{i=1}^{m}P(C_i|R_j)=1$,$P(C_i|R_j)$ 为条件概率。

遍历计算依赖度 $RY_i(R_j)(i=1,2,\cdots,m)$,以原因事件为横轴,以依赖度为纵轴绘制二维图后取截集,实现“由果寻因”的知识推理。

定义5　$\exists C_i\in C(i=1,2,\cdots,m)$,$C_i$ 事件对集合 R 中各元素存在或重或轻(或无)影响,对 $R_j(j=1,2,\cdots,n)$ 而言,C_i 对 R_j 的影响程度记为影响度 $EF_j(C_i)$。

当用 C_i 对集合 R 中各元素的概率因果度平方和的开方值表示影响合力时,有:

$$EF_j(C_i)=\frac{P(R_j|C_i)\sigma_{ij}}{\sqrt{\sum_{i=1}^{n}[P(R_j|C_i)\sigma_{ij}]^2}}\tag{5}$$

其中,$\sum_{i=1}^{n}P(R_j|C_i)=1$,$P(R_j|C_i)$ 为条件概率。

遍历计算影响度 $EF_j(C_i)(j=1,2,\cdots,n)$,以结果事件为横轴,以影响度为纵轴绘制二维图后取截集,实现“由因索果”的知识推理。

3　应用实例

上述方法应用于“嵌入系统总线时序自动分析器”这一研究课题,该课题立足于以人工智能手段实现对种类繁多的嵌入系统总线时序自动分析,为编写接口协议提供决策参考。以分析8输入/3输出的总线信号系统为例,时事折换图见图1。输入/输出信号的事件频度矩阵如下:

$$T|R=\begin{bmatrix}9 & 4 & 6\\ 13 & 16 & 12\end{bmatrix}^{\mathrm{T}}$$

$$T|C=\begin{bmatrix}1 & 2 & 3 & 8 & 9 & 5 & 7 & 4\\ 11 & 6 & 15 & 11 & 14 & 10 & 12 & 6\end{bmatrix}^{\mathrm{T}}$$

因果度定义 k_1 在本例的物理意义是信号沿触发系数,k_2 的物理意义是信号电平触发系数。取 $k_1=k_2=1$,事件因果度计算结果为:

$$\boldsymbol{\sigma}=\begin{bmatrix}0.6250 & 0.9167 & 1.0333\\ 0 & 0.6667 & 0.2500\\ 1.6667 & 1.9167 & 1.8333\\ 1.5000 & 0 & 0\\ 3.2500 & 0 & 0.8333\\ 0.5000 & 0 & 1.6667\\ 1.2500 & 0 & 1.0333\end{bmatrix}$$

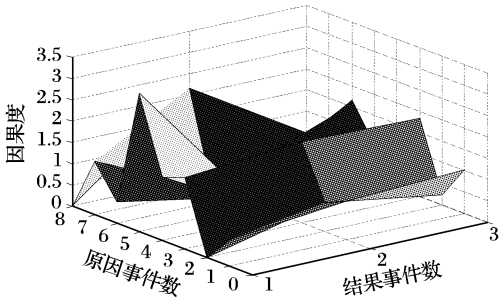

图3　事件因果度三维图

由图3可知,尖峰处对应可能存在因果关系的事件对,低谷处对应不存在因果关系的事件对。例如对 R_1 而言,可能存在因果关系的原因事件集合为 $\{C_1, C_3, C_4, C_5, C_6, C_7\}$,集合 $\{C_2, C_8\}$ 被排除。

在获取上述知识后,继续进行知识推理。设先验概率 $P(C_i|R_j)=1/8$,$P(R_j|C_i)=1/3(i=1,2,\cdots,8;j=1,2,3)$ 事件依赖度和影响度计算结果见图4。

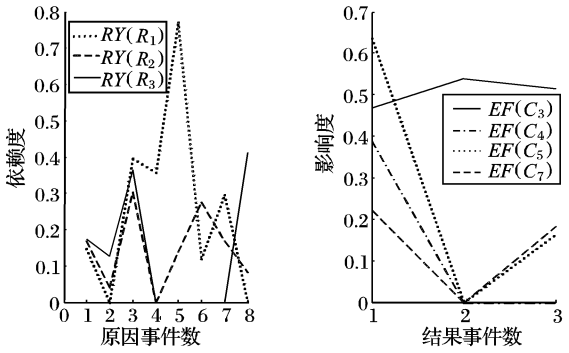

图4　事件依赖度、影响度二维图

在依赖度二维图上,对 R_1 取截集,例如取系数0.3,获取其原因事件为 $\{C_3, C_4, C_5\}$,且有信息 $RY_5(R_1)>RY_3(R_1)>RY_4(R_1)$。在影响度二维图上,对 C_5 取截集,例如取系数0.1,获取其结果事件为 $\{R_1, R_3\}$,且有信息 $EF_1(C_5)>EF_3(C_5)$。同理可进行其余事件的知识推理。

4　结语

协同因果事件族是广泛存在于人类生活的因果事件形态,如何对其进行有效的描述、聚类以及知识推理是涉及众多技术领域的一个基础课题。本文从因果事件的时态数据挖掘入手,运用时事折换思想建立包含时态信息的因果知识事件域描述机制,通过事件因果度、依赖度和影响度量化指标实现因果知识发现及双向推理。该方法具有建模便捷、刻画细致、结果直观等优点,适合高维协同因果事件序列的自动分析推

(下转第2486页)

3 算法的有效性分析

在问题空间中,每一个体是一个排列,对应于编码空间的一个整数值,在进化过程中,每一个体通过高斯变异生成一子个体,生成的子个体是一个实数值,通过取整操作成为有效的个体,式(4)中对 $n!$ 的取模操作是对变异的越界处理。把式(4)改写如下:

$$x_i' = (x_i + v_i) \bmod n! \quad (5)$$

注意到式(5)中对 $n!$ 的取模操作,v_i 相当其值在 $0 \sim n!$ 的自然进位制数,若 v_i 的前 $k(0 \leqslant k < n-1)$ 位均为0,则 k 越大,变异量越大;k 越小,变异量越小。由性质2,x_i 和 v_i 的后 $n-1-k$ 位通过逐位相加运算后的结果有以下两种情况:

1)结果仍为 $n-1-k$ 位,这时再与 x_i 的前 k 位组成 x_i',这时 x_i 与 x_i' 的前 k 位相同。

2)结果为 $n-k$ 位,它的最高位必定为1,把最高位1再与 x_i 的相应位相加,所得结果又有两种情况:无进位,这时 x_i' 与 x_i 的前 $k-1$ 位相同;向第 $k-1$ 位有进位,这同时又可以继续分解下去。再根据性质3,我们可以得出如下结论:

定理1 编码空间上的父个体按式(4)产生子个体,当变异量较大时,父个体和子个体对应于问题空间上的两个全排列的海明距离一定较大;当变异量较小时,两者的海明距离以较大的概率较小。

这一性质符合进化算法求解的思想,算法以较大的变异量实现全局优化,以较小的变异量实现局部优化,从而说明了算法的有效性。

两个体分别为 $n-1$ 位自然进位制数,进行单点交叉后,仍分别为合法的 $n-1$ 位自然进位制数,从而交叉操作不会产生无效的个体。

4 仿真实例

图的最小色数问题是一个典型的NP难问题。给定一个图(如图1所示),使得有边相连的顶点着不同的颜色,要求找到一个最小色数的着色方案。

对顶点进行编号,其编号分别为1,2,…,19。其目标是要求在这19个自然数的全排列中,找出其中一个排列,它把能着同一种色的顶点相邻地排在一起,而且数最小。这也是计算过程中目标函数的计算方法。取群体规模 $u=30$,进化最大代数 $G=30$,$pc=0.4$,得到如下结果。

最小的色数为4,顶点被分成了4类,每一类着同一种颜色,其中:第1类:9,17,10,1,18,11;第2类:12,13,4,6,2,16;第3类:8,19,7;第4类:5,15,14,3。

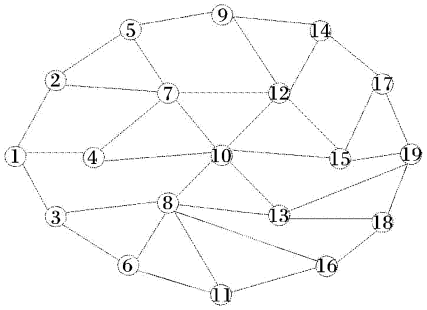

图1 一个无向图

针对上述问题把本文提出的算法与遗传算法相比较。遗传算法采用与文献[7]相同的框架,采用单点交叉映射法,交叉概率取0.4,采用对换变异法,变异率取0.052,两个算法的群体规模与进化最大代数均相同,各为30,两个算法各运行100次。分别求出两种算法在各次运行所得的最小分类数的算术平均。其结果如下:本文算法100次运行所得结果的平均值为4.32,而采用遗传算法运行100次所得结果的平均为4.67,从而说明了本文算法略优于遗传算法。

参考文献:

[1] 李敏强,寇纪淞,林丹,等.遗传算法的基本理论与应用[M].北京:科学出版社,2002.

[2] YAO X. Evolutionary programming made fast[J]. IEEE Transaction on Evolution Computation, 1999, 3(2): 82 - 102.

[3] 梁艳春,冯大鹏,周春光.遗传算法求解旅行商问题时的基因片断保序[J].系统工程理论与实践,2000, 20(4): 7 - 12.

[4] 陶世群,蒲保兴.基于遗传算法的多级目标非平衡指派问题求解[J].系统工程理论与实践,2004,24(8):81 - 85.

[5] LARRANAGA P. Genetic algorithms for the traveling salesman problem: a review of representations and operators [J]. Artificial Intelligence Review, 1999, 13(2): 129 - 170.

[6] 王贵竹,李津生,洪佩琳.变进位记数制与伪随机序列的产生[J].中国科学技术大学学报,2000,30(4):438 - 443.

[7] 蒲保兴,陶世群.遗传算法求解图的染色问题[J].电脑开发与应用,2001,14(2):26 - 27.

(上接第2483页)

理,在信号分析、故障诊断、案情剖析等诸多领域有较大应用潜力。

参考文献:

[1] WOOD J, SULE V, ROGERS E. Causal and Stable Input/Output Structures on Multidimensional Behaviours[C]// 44th IEEE Conference on Decision and Control. Piscataway: IEEE, 2005: 4554 - 4559.

[2] NIKOVSKI D. Constructing Bayesian networks for medical diagnosis from incomplete and partially correct statistics[J]. IEEE Transactions on Knowledge and Data Engineering, 2000, 12(4): 509 - 516.

[3] STERRITT R, MARSHALL A H, SHAPCOTT C M, *et al*. Exploring dynamic Bayesian belief networks for intelligent fault management systems[C]// 2000 IEEE International Conference on Systems, Man, and Cybernetics. [S.l.]: IEEE, 2000: 3646 - 3652.

[4] SHI Q X, WANG H C, ZHANG Q. Intelligent fault diagnosis technique based on causality diagram[C]// Fifth World Congress on Intelligent Control and Automation. Piscataway, NJ: IEEE, 2004: 15 - 19.

[5] WONG M L, LAM W, LEUNG K S, *et al*. Discovering knowledge from medical databases using evolutionary algorithms[J]. IEEE Engineering in Medicine and Biology Magazine, 2000, 19(4): 45 - 55.

[6] 樊兴华,张勤,孙茂松,等.多值因果图的推理算法研究[J].计算机学报, 2003, 26(3): 310 - 322.

[7] WEN F H, LAN Q J, MA C Q, *et al*. An algorithm for mining association patterns between two time series and application in finance [C]// The Sixth World Congress on Intelligent Control and Automation. Piscataway, NJ: IEEE, 2006: 5938 - 5942.

[8] KARIMI K, HAMITON H J. Time Sleuth: a tool for discovering causal and temporal rules[C]//14th IEEE International Conference on Tools with Artificial Intelligence. Montreal: Artificial Intelligence Inc, 2002: 375 - 380.