

基于约束网络的因果关联规则挖掘研究

崔 阳¹ 刘长红²

(中国劳动关系学院数学与计算机部 北京 100048)¹ (江西师范大学计算机信息工程学院 南昌 330022)²

摘 要 因果关联规则是知识库中一类特殊且重要的知识类型,相对一般关联规则,其优势在于能够揭示深层知识。首先对因果关系的特征和因果关联规则的挖掘方法进行了简介。针对在挖掘初始阶段如何限定可能导致结果的原因变量集合这一问题,运用了约束网络原理来构建一个实际系统变量间的因果关系结构。通过该因果关系结构可以比较容易地导出原因变量集合及各变量的类型,从而降低挖掘的复杂性,为提高挖掘结果的准确性提供有利条件。约束网络的引入优化了因果关联规则的挖掘过程,使之趋于更完备。

关键词 因果关联规则,约束网络,因果关系结构,有限表

中图法分类号 TP311.13 文献标识码 A

Research on Causal Association Rule Mining Based on Constraint Network

CUI Yang¹ LIU Chang-hong²

(Mathematics and Computer Department, China Institute of Industrial Relations, Beijing 100048, China)¹

(College of Computer Information Engineering, Jiangxi Normal University, Nanchang 330022, China)²

Abstract Causal association rule is one kind of important and special knowledge type in knowledge base, which can reveal the deep knowledge relative to general association rule. Characteristics of the causal relation and the causal association rule mining were briefly introduced at first. Based on that, constraint network theory was used to construct a causal relation structure of variables in practical system so as to solve the problem of how to restrict the casual variable set in initial step of the mining. Casual variable set as well as types of the variables can be inferred easily from the causal relation structure, so complexity of the mining would be reduced and the results would be more accurate. Introduction of the constraint network can make the process of causal association rule mining become more complete.

Keywords Causal association rule, Constraint network, Causal relation structure, Limited table

1 引言

关联规则是数据挖掘中最为成熟的领域之一。但一般的关联规则也存在不足之处,即规则更多体现的是“联系”而非“因果”。因果关联规则就是指规则的前件与后件之间存在因果关系,由于“因”的出现而导致“果”的发生。因果关联规则具有两个特点:1)规则的前件与后件之间不但具有关联性,而且具有因果性;2)可以使用较为完备的推理机制进行推理^[1]。相对于一般的关联规则,因果关联规则属于深层知识类型,能更好地反应客观事物之间本质和内在的联系。

约束描述了系统中某些给定变量间相互制约、相互依赖的关系,约束网络就是由一些共享某些变量的多个约束组成的系统。在定性推理等领域的研究中,通过说明约束网络中变量关系的因果语义,可以将变量关系转化为因果关系。本文拟通过使用约束网络与因果分析算法相结合的方法生成系统的因果关系结构,明确变量间的因果关系,为因果关联规则的挖掘提供先决条件。

2 因果关系和因果关联规则

2.1 因果关系的含义

因果关系是客观事物普遍联系和相互作用的一种表现形式。因果关系具有因果性、非对称性和可推理性等特征。因果关系大体可分为机制型和时序型两种,前者从原因导致结果的方式和机制上阐述因果关系,后者从过程上表述原因和结果的先后出现关系。

因果关系的前件“原因”主要包括孤立系统某一时刻的存在状态、对某结果的产生起作用的内外因素之全体等;后件“结果”主要包括孤立系统某一(后继)时刻的存在状态、一般环境中的某个原因所产生的全部后效等。基于此,可将因果关系表述为:因果关系是事物在时间上演化发展的法则,本质上是时序与过程的关系,即事物引起与被引起、产生与被产生的关系,具有解释事物或预测变化的功能。

2.2 因果关联规则的特征

在实际情况中,事物间的因果性往往不明显,甚至因被一些表面现象所覆盖导致隐藏较深、结构复杂,而难于为人们所

本文受国家自然科学基金资助项目(61202209)、江西省教育厅科技项目(GJJ13229)资助。

崔 阳(1979—),男,博士,讲师,主要研究方向为知识工程与知识发现,E-mail:cuiyang14@163.com;刘长红(1977—),女,博士,讲师,主要研究方向为图像处理与模式识别。

认识和挖掘。因果关联规则的主要特征可概括为:

- (1) 规则的前件(原因)包含主观原因、偶然原因、间接原因等;规则的后件包含人为结果、偶然结果、意外结果等;
- (2) 规则的前件(原因)状态和后件(结果)状态不完全可知、不可测或常被忽略;
- (3) 规则的前后件相互作用中不但体现了过程特征,而且同时要体现时序特征;
- (4) 规则为链式结构、间接存在,难以显式表达;
- (5) 规则具有不确定性、模糊性特征,对其表示、挖掘和评价要从因和果两方面进行。

通过以上论述可知,因果关系所承载的知识,比关联关系更为全面和深层,同时其结构和挖掘也更为复杂。因果知识可用于故障预测与诊断、专家系统的构建^[2]等。

2.3 基于语言场的因果关联规则挖掘

目前对因果关联规则的研究工作较少,较有代表性的是基于语言场的方法^[3]。语言场及语言值的定义表述如下。

定义 1 在语言变量相应的基础变量论域中,各个被划分的交叉区间的中点连同 ϵ 邻域(ϵ 通常为允许误差值)内的点称为标准样本(点),其取值邻域称为标准值;其余诸点均称为非标准样本(点),其取值称为非标准值。

定义 2 $C = \langle D, I, N, \leq_N \rangle$, 若满足下列条件,则称 C 为语言场:

- (1) D 为基础变量论域 R 上交叉闭区间的集合, $D+$ 为其对应开集;
- (2) $N \neq \emptyset$ 为语言值的有限集;
- (3) $\leq_N$ 为 N 上的全序关系;
- (4) $I: N \rightarrow D$ 为标准值映射,对 $\leq_N$ 满足保序性。

定义 3 对于语言场 $C = \langle D, I, N, \leq_N \rangle$, 称 $F = \langle D, W, K \rangle$ 为 C 的语言值结构,如果:

- (1) C 满足定义 2;
- (2) K 为自然数;
- (3) $W: N \rightarrow R^k$ 满足:

$$\forall n_1, n_2 \in N (n_1 \leq_N n_2 \rightarrow W(n_1) \leq_{dic} W(n_2))$$

$$\forall n_1, n_2 \in N (n_1 \neq n_2 \rightarrow W(n_1) \neq W(n_2))$$

其中, $\leq_{dic}$ 为 R^k 上的字典序,即 $(a^1, \dots, a^k) \leq_{dic} (b^1, \dots, b^k)$ 当且仅当存在 h , 使得当 $0 \leq j < h$ 时 $a^j = b^j, a^h \leq b^h$ 。

基于以上定义,将因果关联规则表示为 $[A_i] \rightarrow [S_i]$ 的形式,其中 $[A_i]$ 与 $[S_i]$ 分别表示原因与结果所处(变)态的语言值形式,“ $\rightarrow$ ”表示因果关联关系。在此基础上给出了基于广义归纳逻辑因果模型^[3]的因果关联规则挖掘算法,该算法可以完整地表示、挖掘和评价一个给定系统中的因果关联规则,并且能够较好地刻画出因果关联规则的时序特征。

3 使用约束网络发现因果关系结构

从上述因果关联规则挖掘算法可知,在挖掘开始前对导致结果的原因变量及变量的类型(直接、间接等)进行限定是一个重要问题。通过领域专家直接给出原因变量集合的方式不但易受主观因素影响,而且面对复杂系统时也比较困难。考虑到一般情况下一个实际系统可以用一些数学方程组描述,而这些方程组中就蕴含含有变量之间的因果关系,可以采用约束网络方法来对方程组求解。

3.1 约束网络中的因果性

在约束描述系统中给定变量的关系,多个约束组成一个系统的约束网络。约束网络常用于表示和分析因果关系。例如,在基于约束的定性物理方法中,因果性表现为扰动作用在约束方程中的传播,也称虚构的因果性^[4]。

在因果顺序理论中,通过了解各变量出现在哪个方程中,从而求解自容方程组中变量之间的因果关系^[5]。近年还出现了因果时间约束网络模型^[6],这种模型通过建立定性、定量的时间约束和引入因果约束等方法,在不同的解释情景中构建信息,并用一个获取与初始时间信息相一致的解释的推导机制将这些信息彼此连接在一起,实现时间信息的表示和因果关系的处理。

约束网络通常没有方向性,这是因为约束只说明集合中哪些值是相容的,而相容性是对称关系,但因果关系则是非对称的。因此对约束网络中的边还要给出方向性的关系语义,从而将无方向的约束网络转化为有方向的因果结构。

3.2 求解因果关系结构

首先给出约束的严格定义:称变量 $x_0, x_1, \dots, x_n$ 中存在一个约束,是指在 $x_0, x_1, \dots, x_n$ 中任意确定 n 个变量,则剩下的一个变量也随之确定。由 n 个约束组成的集合就称为约束集,记为 S 。

从因果关联规则挖掘的实际需求出发,采用因果顺序方法求解系统的因果关系结构。其大体思路是首先将表示系统的数学方程组转化为约束方程组,形成约束网络;然后采用因果分析算法将约束网络转化成果因果关系结构。

为求解系统的因果结构,首先要将描述系统的方程组、变量间已知的定性关系等转化为约束集。方法是去掉原方程组中各方程的运算,将变量间的定性关系进行抽象,形成变量间的约束关系;然后将这些约束关系转化为选定的约束形式,以进行因果分析。这里选用有限表作为约束的数据结构。表中每个元素代表系统一个变量,每行表示一个约束^[7]。

以某一实际系统为例:该系统中变量有 A, B, C, D, E, F, G 和 H ,假设变量间的关系可用如式(1)~式(4)表示:

$$E = k(H - G) \quad (1)$$

$$E = lDA - m(B - C)A \quad (2)$$

$$G = nF \quad (3)$$

$$C = G \quad (4)$$

其中, k, l, m 和 n 为常量系数。如果再有一组初值方程:

$$A = a \quad (5)$$

$$B = b \quad (6)$$

$$F = f \quad (7)$$

$$H = h \quad (8)$$

则式(1)~式(8)构成了描述系统的方程组。将这个方程组中的运算去掉,就得到了系统的约束方程组:

$$(E \ H \ G) \quad (9)$$

$$(E \ D \ A \ B \ C) \quad (10)$$

$$(G \ F) \quad (11)$$

$$(C \ G) \quad (12)$$

$$(A) \quad (13)$$

$$(B) \quad (14)$$

$$(F) \quad (15)$$

$$(H) \quad (16)$$

接下来将约束方程组转化为有限表。表 1 列出了方程组的结构,每一行代表一个方程。 T 表示相应的位置上有变量出现, F 表示不出现。

表 1 系统变量的有限表

	A	B	C	D	E	F	G	H	H
(9)	F	F	F	F	T	F	T	T	T
(10)	T	T	T	T	T	F	F	F	F
(11)	F	F	F	F	F	T	T	F	F
(12)	F	F	T	F	F	F	T	F	F
(13)	T	F	F	F	F	F	F	F	F
(14)	F	T	F	F	F	F	F	F	F
(15)	F	F	F	F	F	T	F	F	F
(16)	F	F	F	F	F	F	F	T	T

在得到有限表后对其进行因果分析,以便从约束网络中找出系统的因果结构。这里使用有限表因果分析算法完成这一工作。该算法的思路是给定一个自容约束集 S , 设 S_0 为所有最小完备子集的并集。最小完备子集是指如果存在自容约束集 S 的自容子集 s 且 s 中没有自容子集, 则称 s 为一个最小完备子集。将 S_0 的解代入 $(S - S_0)$ 的方程, 得到新的自容结构。重复上述过程直到导出结构中不再含有自容的子集。对 S 中的每个方程 e_i , 令 V_i 表示 e_i 中出现的变量集合, W_i 为 V_i 中属于最高阶的最小完备子集的变量集合, 则 W_i 中的变量称为直接因果依赖于 $(V_i - W_i)$ 的变量。依据此从表 1 中得到 W_i, V_i 和 $(V_i - W_i)$, 如表 2 所列。

表 2 系统变量的直接因果依赖关系

V_i	W_i	$(V_i - W_i)$	直接因果依赖
$\{E, G, H\}$	$\{E\}$	$\{G, H\}$	$\{E G, H\}$
$\{A, B, C, D, E\}$	$\{D\}$	$\{A, B, C, E\}$	$\{D A, B, C, E\}$
$\{F, G\}$	$\{G\}$	$\{F\}$	$\{G F\}$
$\{C, G\}$	$\{C\}$	$\{G\}$	$\{C G\}$
$\{A\}$	$\{A\}$	$\emptyset$	$\emptyset$
$\{B\}$	$\{B\}$	$\emptyset$	$\emptyset$
$\{F\}$	$\{F\}$	$\emptyset$	$\emptyset$
$\{H\}$	$\{H\}$	$\emptyset$	$\emptyset$

由此就得到了系统中各变量间的因果依赖关系。根据表 2 可以绘制出变量间的因果关系结构图, 如图 1 所示。

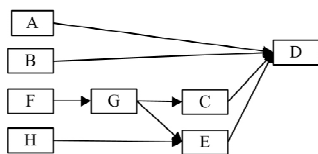


图 1 系统变量的因果关系结构图

以变量 D 为例, 从图 1 中可以看到, 导致结果 D 的直接原因变量有 A, B, C 和 E , 而间接原因变量有 G, F 和 H 。再如, 导致结果 E 的直接原因变量有 G 和 H , 间接原因变量为 F 。这样就在因果关联规则挖掘过程开始前明确了原因和结果变量的关系。

由约束网络转化而来的因果关系结构图的优点之一是在系统结构不变的情况下, 如果变量间约束关系有所改变, 或用户关注重点发生转移, 则只须根据因果关系结构图使用特定

算法对其稍作转换, 即可得到新的因果关系结构图。

3.3 因果关系语义解释

如前所述, 约束网络没有方向性, 因此需要为其添加具有方向性的语义, 以表示非对称的因果关系。完整准确的因果语义形式可用于对系统进行解释或诊断。分析因果关联规则的含义, 可知在因果关系中主要关注的是原因和结果的类型、时间特征、影响强度等。因此在因果语义中应体现这些内容。给出如图 2 所示的因果语义表示。

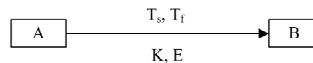


图 2 因果语义图

其中, (T_s, T_f) 表示因果影响的持续时间; K 表示因果类型; E 表示因果强度, 可以是时间的函数。因果语义具有重要作用。一方面, 因果语义使对称、无方向的约束集合转化为有方向性的因果结构; 另一方面, 通过因果语义, 可以实现对因果结构的解释、诊断、推理。特别是 Dechter 和 Pearl 等人通过给出这种方向性的因果语义得到一个重要结论: 预测比诊断和规划更容易实现^[8]。因此, 在故障发生前进行预测较在故障发生后进行诊断, 无论从时间还是从成本上都更具有优势。这也是因果关系的重要应用之一。

结束语 因果关联规则可以反映深层知识。将约束网络方法用于因果关联规则的挖掘过程, 通过转化约束网络得到系统的因果关系结构, 找出造成结果的各原因变量, 能有效地为因果关联规则挖掘过程提供先决条件。同时约束网络方法也存在处理循环反馈情况的因果关系的能力较弱等不足, 今后工作将就此类问题展开进一步研究。

参 考 文 献

- [1] 杨炳儒. 基于内在机理的知识发现理论及其应用[M]. 北京: 电子工业出版社, 2004: 152-153
- [2] Iwasaki Y, Simon H A. Theories of causal ordering: Reply to de Kleer and Brown[J]. Artificial Intelligence, 1986, 29(1): 63-72
- [3] 杨炳儒, 钱榕, 张伟. 语言场理论及其在知识发现中的应用[J]. 计算机工程, 2005, 31(24): 10-12
- [4] De Kleer J, Bronw J S. A qualitative based on confluences[J]. Artificial Intelligence, 1984, 24(1-3): 7-83
- [5] Iwasaki Y, Simon H A. Retrospective on causality in device behavior[J]. Artificial Intelligence, 1993, 59(1): 141-146
- [6] Angel F L, Vicente M B, Eduardo M R. Causal temporal constraint networks for representing temporal knowledge[J]. Expert Systems with Applications, 2009, 36(1): 27-42
- [7] 石纯一, 廖士中. 定性推理方法[M]. 北京: 清华大学出版社, 2002: 122-124
- [8] Verma T, Pearl J. Causal Networks: Semantics and Expressiveness[J]. Machine Intelligence & Pattern Recognition, 2013, 9: 69-76