知识发现中的因果关联规则挖掘研究

崔 阳,杨炳儒
CUI Yang,YANG Bing-ru

北京科技大学 信息工程学院,北京 100083
Department of Computer Science and Engineering,University of Science and Technology Beijing,Beijing 100083,China
E-mail:cuiyang14@163.com

CUI Yang,YANG Bing-ru.Research on causal association rule mining in knowledge discovery.Computer Engineering and Applications,2009,45(31):9-11.

Abstract: Causal association rule is one kind of important and available knowledge in knowledge base.In this paper,some special properties of causal relation are discussed at first,and then research on represention,mining and evaluation of causal association rule is introduced in detail based on language field and generalized inductive logic causal model.Finally,a new concept of hidden causal association rule is prospected.The application of language field and reasoning mechanism make the process of mining and evaluation of causal association rule logical and extensible.

Key words: causal association rule;language field;knowledge discovery;hidden causal association rule;hidden causal association rule

摘 要: 因果关联规则是知识库中一类重要的知识类型,具有重要的应用价值。首先对因果关系的特殊性质进行了分析,然后基于语言场和广义归纳逻辑因果模型,从表示、挖掘、评价和应用几方面,对因果关联规则的研究进行了详细论述。并在此基础上提出了隐含因果关联规则的概念。通过语言场和推理机制的运用,使因果关联规则这一重要知识形式的挖掘和评价过程具有良好的逻辑性和扩张性。

关键词: 因果关联规则;语言场;知识发现;隐含因果关联规则
DOI: 10.3778/j.issn.1002-8331.2009.31.003 **文章编号:** 1002-8331(2009)31-0009-03 **文献标识码:** A **中图分类号:** TP311.13

1 引言

关联规则挖掘一直是数据挖掘领域中最重要、最活跃的研究内容之一。关联规则描述事务之间的关联性,但是通过支持度和置信度这两个评价标准可知,关联规则的挖掘过程中仅仅关注联合概率表的一部分,因此存在着无法完全反映出事务间相关性的缺点[1]。针对这一问题,通常采取一些关联分析方法,对得到的关联规则挖掘结果作进一步分析,以便尽可能多地获取有意义的和用户感兴趣的规则,但尚不能完全克服该缺点。

因果关联规则是一类特殊的关联规则,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。因果关联规则与一般关联规则的不同之处在于:规则的前件与后件之间不但具有关联性,而且具有因果性;可以使用较为完备的推理机制进行推理[2]。有关因果关联规则的挖掘,目前专门的研究还不是非常多。但实际上因果关联规则是知识发现中一个重要的知识类型,它能够反应客观事物之间更为本质和内在的联系。因此该文拟从表示、挖掘、评价几方面对因果关联规则进行研究。

2 因果关系与关联关系的比较

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,具有普遍的意义。相对于关联关系,因果关系具有以下一些特殊性质:

(1)非对称性。关联关系和因果关系间最显著的不同在于关联是对称的,而因果关系是非对称的。如果两个事务具有因果性,则表示其中一个事务对另外一个事务具有某种影响力,反之不然。有因必有果,因果关系必然存在着关联性;但如果两者存在关联性,则未必一定有因果关系[3]。这说明因果关系成立的条件要比关联规则高。

(2)因果性。对于形如 A→B 的关联关系,可知 A 与 B 之间存在关联性,但无法获取 A 对 B 施加影响的结果;而因果关系不但能够指出 A 与 B 间的关联性,更可以给出 A 对 B 的影响结果。造成这种差异的原因是关联关系仅仅是一种表象上的联系,而因果关系反映的则是事务内在机制性的联系。这说明,根据关联关系,可以使用两事务中的任意一个对另一个进行预测;但如果要对其中一个事务加以改变,并预测会对另一个事务产生何种影响,就需要根据二者之间的因果关系来观察。

(3)可推理性。通常可以使用因果关系图来描述一个系统内各事务之间的因果关系。这种因果关系图是有向图。图的结点表示事务,有向边则用来表示非对称的因果关系。一个基本的因果关系如图 1 所示:

基金项目: 国家自然科学基金(the National Natural Science Foundation of China under Grant No.60675030)。
作者简介: 崔阳(1979-),博士生,主要研究方向:知识发现;杨炳儒(1943-),教授,博士生导师,主要研究方向:知识工程与知识发现,柔性建模。
收稿日期: 2009-08-21 **修回日期:** 2009-09-27

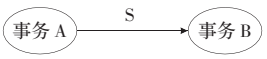

图 1 因果关系的基本结构

其中 S 表示因果语义,包括时间、因果关系类型、影响效果等多种。因果语义非常重要,因为因果关系必须与因果语义相结合才能进行推理。由于语义的不同可能会导致对同一因果关系产生不同甚至相反的解释,所以在进行因果推理时要明确因果关系的语义内容。因果路径及传播过程是因果推理的结果。

因果关系的这种可推理性具有重要的实用价值。例如,在故障诊断领域的发展中存在两种趋势,一种是使故障诊断向智能化方向发展,另一种则将重点放在事先预测方面。后一种方法的思路是通过为物理系统建立因果模型,并给出详细的因果语义。通过因果模型分析物理系统的行为,从而预测可能出现的故障。其根据是从 Dechter 和 Pearl 等人通过对因果关系的研究而得到的重要结论:预测比诊断和规划更容易实现[4]。而且,在故障发生前进行预测,较之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。

3 因果关联规则的研究

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。在基于数据库的知识发现(KDD)中,所发现的知识的主要形式是规则。因此,因果关联规则很自然成为比较重要的一种知识类型。但较之关联规则,对因果关系和因果关联规则的研究工作尚不多见。

以下主要从因果关联规则的表示、挖掘、评价几方面对其进行研究。

3.1 因果关联规则的表示

知识的表示方式多达三十余种。这些知识表示方式各有所长,也都有其针对性和局限性。规则是知识库中知识的主要存在形式[5]。选择基于语言场的知识表示方法来表示因果关联规则。

语言场及语言值结构的相关定义如下[6]:

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。它们分别构成标准样本空间与非标准样本空间,并统称为一般样本空间。

定义 2 $C=<D,I,N,\leq_N>$,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)$N\neq\phi$ 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,满足保序性,即:$\forall n_1,n_2\in N(n_1\neq n_2\wedge n_1\leq_N n_2\rightarrow I(n_1)\leq I(n_2))$($\leq$为偏序关系)。

定义 3 对于语言场 $C=<D,I,N,\leq_N>$,称 $F=<D,W,K>$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$(a^1,a^2,\cdots,a^k)\leq_{dic}(b^1,b^2,\cdots,b^k)$当且仅当存在 h,使得当 $0\leqslant j<h$ 时 $a^j=b^j$,$a^h\leqslant b^h$。

定义 4 表达因果关系的产生式规则称为因果关联规则,其一般形式为:

$[A_i]\hookrightarrow[S_i]$

其中$[A_i]$与$[S_i]$分别表示原因与结果所处状(变)态的语言值形式;"$\hookrightarrow$"表示因果关联关系。多原因的情形依此类推。

3.2 因果关联规则的挖掘

提出的因果关联规则的挖掘算法基于广义归纳逻辑因果模型。之所以选择广义归纳逻辑因果模型,主要是考虑到其具备较好的归纳推理机制,有利于在此基础上形成不确定性因果归纳推理的计算模型和自动推理机制[7]。

如果语义结构 $X^*=\langle S^*,\Pi^*\rangle$满足:

(1)$S^*=\{S_a,S_1,S_2,\cdots,S_M\}$,$S_a$为现实的因果世界;$S_i(i=1,2,\cdots,M)$ 为首因果必然性规律与有关→的原理所支配的可能因果世界;$S_i=\{V_{i1},V_{i2},\cdots\}$,$V_{ij}$表示组成 S_i的不同历史,每个历史含有不同的时空段,每个时空里潜含着各类因果联系,而因与果又对应着各自的语言场与语言值结构。

(2)Π^* 是广义因果细胞自动机,每个可能的因果世界都用相应的广义因果细胞自动机来描述。

则 X^* 为广义归纳逻辑因果模型。

算法将因果关联规则的挖掘分为标准样本空间和一般样本空间两类。

标准样本空间中,当用广义因果细胞自动机去描述标准样本空间在时刻 t 的因果间的状(变)态联系时,首先得到因果各种状(变)态的语言值描述及其对应的离散型向量表示。

定义 5 在标准样本空间中设 $\boldsymbol{A}_t^{(i)}$与 $\boldsymbol{S}_t^{(j)}$分别表示原因 A 在 t 时刻状(变)态与结果 S 在 t'时刻状(变)态的标准向量。则因果状态必然性规律 $\varphi_i^*(A,t)\rightarrow\varphi_j^*(S,t')$由带有模糊性与随机性的状态的关系矩阵给出,即:

$$\varphi_i^*(A,t)\rightarrow\varphi_j^*(S,t')\triangleq C(H,E)\cdot[(\boldsymbol{A}_t^{(i)})^{\mathrm{T}}\times\boldsymbol{S}_{t'}^{(j)}]$$

其中,$C(H,E)$为归纳确认度函数,它表明证据 E 对假说 H(即此表示因果状态的必然性规律)的支持程度。假说 H 的归纳确认度函数是两个测度矩阵乘法的范数之比:$C(H,E)=\|SE\|/\|AE\|$。

根据广义细胞自动机的定义中的因果状(变)态原理,结果 S 在时刻 t'的所有可能状(变)态可由下式获得:

$$(\varphi_i^*(A,t)\hookrightarrow\varphi_j^*(S,t'))\wedge(\varphi_j^*(A,t'')\rightarrow\varphi_k^*(S,t'''))=C(H,E)\cdot[(\boldsymbol{A}_t^{(i)})^{\mathrm{T}}\times\boldsymbol{S}_{t'}^{(j)}]+C'(H,E)\cdot[(1-\boldsymbol{A}_{t''}^{(i)})^{\mathrm{T}}\times\boldsymbol{S}_{t'''}^{(k)}]+\cdots$$

其中 $C(H,E)$和 $C'(H,E)$均为归纳确认度函数。

上式中由 $\varphi_i^*(A,t)$所导致的 $\varphi_j^*(S,t')$融会了所有可能的多个结果(标准向量),实现了在标准样本空间中在认知极限内的完全归纳。由上式形成的每一个矩阵均称为状(变)态矩阵,记为 $\boldsymbol{M}$。

在广义归纳逻辑因果模型构造下,在可能因果世界中,以含有因果联系信息的状(变)态矩阵 $\boldsymbol{M}$ 为背景(大前提),要获得原因 A 在某个状(变)态(小前提)下所能导致结果 S 的状(变)态(结论),其归纳推理模式为:

$M_0:\varphi_i^{(0)}(A,t)$　$\varphi_i^{(0)}(S,t)$

$M_1:\varphi_i^{(0)}(A,t)$　$\varphi_k^{(1)}(S,s')$

…　归纳大前提(矩阵集)

$M_n:\overline{\varphi_i^{(0)}}(A,t)$　$\varphi_l^{(n)}(S,t')$

A_t^*　归纳小前提(因向量)

$S\triangleq A_t^*\cdot(M_0+M_1)$ $(j=1,2,\cdots,n)$　归纳结论(果向量集)

从而得到因果关联规则 $A_t^*\multimap\!\!\rightarrow S$。

3.3　因果关联规则的评价

如何对挖掘到的关联规则进行评价，关系到输出规则的数量和质量。目前关于关联规则的评价方法研究多集中在客观感兴趣度的研究，如 Piatesket-Shapiro 提出了事件独立性方法[8]、Symth 提出了 J-Measure 函数[9]等。这些方法共同的缺点是只利用规则的前件和后件的客观关联来评价对规则的感兴趣程度，忽视了背景知识和用户的参与。

相比之下，因果关联规则由于规则的前件与后件之间存在因果联系，因此是一种强关系。因此可以运用因果关系定性推理机制进行推理，并可利用认证逻辑的分析方法实现后验评价的方法[10]。这是因果关联规则在评价方面与一般关联规则的不同之处。

首先，在给定原因的样本值情况下，由因果关系自动推理机制推出果状(变)态；构建评价知识库，作为所有原因为 A 的状(变)态和结果为的 S 的状(变)态规则的评价依据；取样本中原因 A 和结果 S 的数据，构成一个序偶的集合 $P=\{<T_w,S_w>\}$ $(w=1,2,\cdots,N)$，T_w 为原因状(变)态空间中的数据(即因样本值)，S_w 为与原因数据相对应的结果状(变)态空间中的数据(即果样本值)，N 为集合中样本的个数。

将所发现的因果关联规则记为 $R(A_i\multimap\!\!\rightarrow S_j)$，对规则进行评价就是判定是否接受此规则，因此它属于认证逻辑的范畴。评价的关键在于确定验前置信度和验后置信度。

定义 6　对因果关联规则 R $(A_i\multimap\!\!\rightarrow S_j)$，$A_i$ 与 S_j 两者同时出现的概率与两者析取出现的概率之比，即 $P(A_i\wedge S_j)/P(A_i\vee S_j)$ 称为因果关联强度，记作 CR，将其作为验前置信度。

在因果关联规则的评价过程中应用认证逻辑分析方法，需要用到一致性原理和适用性原理。根据适用性原理，假设 R 表示需要评价的因果关联规则，E 为可从 R 推出的一些检验结果集合。在评价过程中，根据因果关系自动推理机制，检验因果数据是否满足一致性原理，即如果样本中果数据的状(变)态等于由原因数据经推理所得的结果，则表明它满足一致性原理；反之不满足。设 $E1$ 是满足一致性原理的检验结果，如果所采用的样本总数为 N，产生这个结果的样本数记为 $N(E1)$；$E2$ 为不满足一致性原理的检验结果组成的集合，产生这个结果的样本数记为 $N(E2)$，必然有 $N(E1)+N(E2)=N$。

定义 7　将 $N(E1)/(N(E1)+N(E2))$ 称为支持强度，记作 SUP。

结论　对于因果关联规则 $R(A_i\multimap\!\!\rightarrow S_j)$，若 $SUP>CR$，则此因果关联规则得到认证；若 $SUP\leq CR$，则此因果关联规则被拒绝。

4　隐含因果关联规则的提出

因果关联规则能够体现数据库中具有因果性的知识。但在实际情况中，事务间的因果性往往不明显，甚至由于为一些表面现象所覆盖导致隐藏较深、结构复杂，而难于为人们所挖掘和认识，更无法加以利用。在对因果关联规则进行研究时，希望找出因果关联规则的一些本质特性，从而深化这一概念，并将这些性质进一步用于因果关联规则的挖掘。这就产生了隐含因果关联规则的概念。

隐含因果关联规则(或称隐含因果关系)，是一类特殊的因果关联规则。通过查阅国内外文献可以发现，目前学术界对这一概念还没有形成完整的定义。我们试图以因果关系和因果关联规则的研究为基础，首先为隐含因果关联规则形成一个较为明确的定义。“隐含”规则的含义包括：

(1)规则的前件(原因)状态和后件(结果)状态尚不完全可知、不可测，或为人所忽略；

(2)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，这些原因和结果很少或从不在因果关联规则中体现；

(3)规则的前后件相互作用中不但体现了过程特征，而且同时体现时序特征；

(4)规则可能是链式结构的、间接存在的；或使用当前的因果关联规则挖掘算法难以挖掘；或规则难以显式表达等；

(5)规则具有不确定性、模糊性特征，对其定义、表示、挖掘和评价应以因果关联规则为基础，从因和果两方面进行。

隐含因果关联规则所包含的原因和结果更为多样化，其联系更为复杂和深层。不确定性、模糊性仍然是隐含因果关联规则的重要特征，且对不同的主体而言，“隐含”的含义也不尽相同。

目前对隐含因果关联规则的研究主要还是在现有因果关联规则挖掘方法上进行改进。例如，在模糊状态描述标准结构 C 中，当 R 定为[0,1]时，则定义 2 和定义 3 将分别变为模糊语言场和模糊语言值结构的定义。这种模糊定义可以用来表示隐含因果关联规则中模糊性和不确定性知识。

5　展望

因果关联规则是一类特殊的关联关系。较之一般关联规则，因果关联规则所具有的因果性等特征使其能够提供更为深层的知识。因果关联规则在故障诊断等领域的实际应用是一个值得进一步研究的问题。隐含因果关联规则是对因果关联规则的扩展和深化，今后的工作重点将从现有研究基础出发，对隐含因果关联规则的挖掘、评价及应用等问题作深入研究。

参考文献：

[1] 贺炜，潘泉，陈玉春，等.关联规则挖掘与因果关系发现的比较研究[J].模式识别与人工智能，2005，18(3)：328-332.

[2] 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.

[3] 文凤.因果关系的表达与逻辑推理[D/OL].中国优秀硕士学位论文全文数据库，2004.

[4] Verma T，Pearl J.Causal networks：Semantics and expressiveness[C]// the 4th Workshop on Uncertainty in Artificial Intelligence，1988：352-359.

[5] 蔡勇，石纯一.定性推理中的一种因果分析方法[J].模式识别与人工智能，1995，8(3)：203-209.

(下转 14 页)