KDD 中因果关联规则的评价方法*

杨炳儒, 綦艳霞

(北京科技大学 信息工程学院,北京 100083)
E-mail: bryang@public.fhnet.cn.net
http://www.ustb.edu.cn

摘要: 在 KDD(knowledge discovery in database)中,对所发现的知识进行评价是一个很重要的环节.提出了一种针对 KDD 中因果关联规则的自动评价方法.该评价方法采用了全新的、有效的知识表示方法(语言场和语言值结构)和推理机制(因果关系定性推理机制),并且具有通用性和交互性的特征.给出了此评价方法的理论依据和构造过程,并提供了相应的算法.通过对具体实例的运行检验,证明了此评价方法的有效性.通过与相关工作的比较,证明了其先进性.

关 键 词: 因果关联规则;感兴趣度;评价;KDD(knowledge discovery in database)
中图法分类号: TP181 文献标识码: A

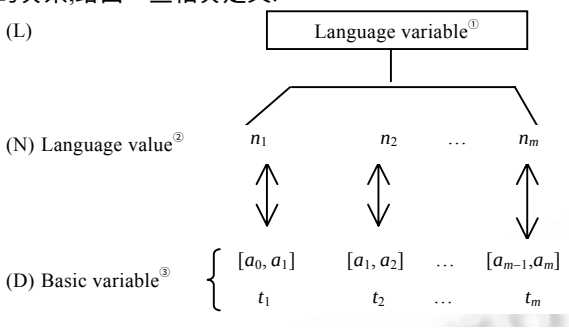
在数据库上的知识发现(knowledge discovery in database,简称 KDD)中,对所发现的知识进行评价是一个很重要的环节,它直接影响着知识发现系统输出的数量和质量.目前,KDD 的主流是关于数据发掘方法的研究,对于评价方法的专门研究很少.规则的感兴趣度是有关所获得规则的有效性、潜在有用性、新颖性和可理解性的综合度量.它根据驱动方式的不同,分为对规则的客观感兴趣度(数据驱动)和主观的感兴趣度(用户驱动).目前关于评价方法的研究比较少,其中主要是对规则的客观感兴趣度的研究.例如,对于规则 $A \Rightarrow B$,Agrawal[1]提出了置信度 $P(B/A)$、Piatesket-Shapiro[2]提出了事件独立性 $P(A,B)/P(A)P(B)$、Symth[3]提出了 J-Measure 函数等.Toivonen[4]提出了根据规则的后件,对挖掘出的关联规则集合进行分组的覆盖集合(cover rules)作为感兴趣的规则.这些方法共同的缺点是,只是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.

在 KDD 中,所发现的知识的主要形式是规则,其中因果关联规则是比较重要的一种知识类型,它在控制、管理等领域都有广泛的应用.但是,目前尚没有确定用于因果关联规则的评价标准.同时,从认知角度讲,对于所获得的知识(假设)的评价有两个步骤:(1) 先验评价.即依据假设发现时知识储备中的证据和公设对假设进行评价;(2) 后验评价.即在知识储备的基础上,加上该假设推出可检验陈述,依据检验结果对假设进行评价.先验评价和后验评价构成了一个完善的评价机制.无论是客观感兴趣度还是主观感兴趣度都属于对规则的先验评价(进一步的分析见第 6 节),都没有通过计算机程序,按推理机制对规则进行后验评价,这主要是因为一般的规则尚不存在比较完备的推理机制.但对因果关联规则这种强的关系而言,在采用了语言场理论表示之后,便可运用已被证实为完备的因果关系定性推理机制进行推理,利用认证逻辑的分析方法,从而实现了后验评价.在这种背景下,本文针对因果关联规则提出了一种全新的评价方法.因为因果关系是普遍存在的,所以此评价方法可以在不同的领域中得到应用,具有通用性.同时,对于有些经常变化的数值,可以通过人机交互,让用户随时进行补充和修改.不过,当所需要的值确定之后,推理机制和评价方法的实现都是由计算机自动完成的,下面分别进行介绍.

* 收稿日期: 2001-02-15; 修改日期: 2001-06-05
基金项目: 国家自然科学基金资助项目(69835001);国家教育部科技重点项目([2000]175)
作者简介: 杨炳儒(1943－),男,天津人,教授,博士生导师,主要研究领域为推理机制与知识发现,柔性建模与系统集成;綦艳霞(1973－),女,山东东营人,博士,工程师,主要研究领域为数据挖掘,人工智能,知识评价.

1 语言场和语言值的知识表示方法

我们先根据如图 1 所示的关系,给出一些相关定义.



①语言变量,②语言值,③基础变量.

Fig.1 Language field and language value structure

图 1 语言场和语言值结构

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.** $C=\langle D,I,N,\leq_N\rangle$,若满足下列条件:

(1) D 为 R 上交叉闭区间的集合(基础变量论域);

(2) $N\neq\varnothing$为语言值的有限集;

(3) $\leq_N$为 N 上的全序关系;

$I:N\rightarrow D$ 为标准值映射,满足保序性,则称 C 为语言场.

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,如果满足以下条件:

(1) C 满足定义 2;

(2) K 为自然数;

(3) $W:N\rightarrow R^k$ 满足:$\forall n_1,n_2\in N(n_1\leq_N n_2\rightarrow W(n_1)\leq_{\mathrm{dic}}W(n2)),\forall n_1,n_2\in N(n_1\neq n_2\rightarrow W(n_1)\neq W(n_2))$.

其中,$\leq_{\mathrm{dic}}$为$[0,1]^k$上的字典序,即$(a^1,\ldots,a^k)\leq_{\mathrm{dic}}(b^1,\ldots,b^k)$当且仅当存在 h,使得当 $0\leq j<h$ 时 $a^j=b^j,a^h\leq b^h$(相关定义及定理详见文献[5]).

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$[A_i]\multimap[S_i]$,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;"$\multimap$"表示因果关联关系.对于多原因的情形依此类推.

在语言场理论中,上述的原因和结果都代表某一语言变量,其所处的状(变)态为相应的某一语言值,并可以根据语言值结构将用自然语言描述的状态离散地表示为论域上的向量.这样,在语言场和语言值结构的描述架下,可以使人们用自然语言的定性表达与可被计算机处理的定量表示之间转化的难题,比较完善地得以解决.

2 因果关系自动推理机制的应用

2.1 因果关系自动推理机制

在给定原因的样本值的情况下,由因果关系自动推理机制推出果状(变)态的过程:

(1) 在一般样本空间中,对于原因 A 而言,其对应的因状(变)态输入向量 a_t,可根据相邻因状(变)态标准向量利用插值公式而获得,即

$$a_t=A_i\cdot\left(1-\frac{|t_i-t_{i0}|}{l_i}\right)+A_{\text{邻}}\cdot\frac{|t_i-t_{i0}|}{l_i}. \tag{1}$$

其中 t_i 为落在第 i 个区间的输入数据,t_{i0} 为第 i 个区间的中点数据,l_i 为第 i 个区间的长度,A_i 为第 i 个区间中的因状(变)态标准向量,$A_{邻}$ 为依 t 的落点而定的左邻或右邻区间中的因状(变)态标准向量.

(2) 给出定义 5.

定义 **5.** 在广义归纳逻辑因果模型中,同一语言值结构下,因状(变)态输入向量 a_t 与因状(变)态标准向量 $A_i^{(j)}$ 的测度由式(2)来确定:

$$d_H(a_t, A_i) = \sum_{j=1}^{n} \left| \mu a_t^{(j)} - \mu A_i^{(j)} \right|, \tag{2}$$

其中 $\mu a_t^{(j)}$ 与 $\mu A_i^{(j)}$ 分别为其各自对应的坐标.根据定义 5,对于原因 A 而言,计算 a_t 与 A 的各状(变)态标准向量的测度,取最小者以确定归属的因状(变)态类型(语言值).

(3) 在因果关系定性推理模型构造下,在可能的因果世界中,根据判定的因状(变)态输入向量 a_t 所属因状(变)态类型(如 $A_t^{(w)}$ 型)以及认定的局部大前提类型(如 $A_t^{(1)} \multimap\!\!\blacktriangleright S_{t'}^{(1)}$),可以在评价知识库(其构造过程见下)中通过自组织的方式找到与其相匹配的惟一的知识矩阵(M_σ^*),以 M_σ^* 为背景(大前提),要获取原因 A 所能导致的结果 S 的状变态(结论),其自动推理模式为

$$\frac{\begin{array}{ll} M_\sigma^* & (\text{大前提}) \\ a_t & (\text{小前提}) \end{array}}{S^* \triangleq a_t \circ M_\sigma^*}. \tag{3}$$

(4) 聚类.确定 S^*归属的结果状(变)态类型(语言值),从而完成了不确定性因果归纳自动推理的全过程.

2.2 评价知识库的构建

(1) 根据对应关系,在文献[6]中给出了因果关联规则在局部大前提为 $A_i \multimap\!\!\blacktriangleright S_j$,$A_i$ 为 A_P 时的因果状(变)态表,同样可以得 A_P 到剩余的 4 个因果变态表(A_i 分别为 A_R,A_T,A_Q,A_S 时的因果状(变)态表).例如:

Table 1 Causal changing state table

表 **1** 因果变态表

Major premise①		Small premise②	Result vector③
$A_P \multimap\!\!\blacktriangleright$	$\overline{A}_P \multimap\!\!\blacktriangleright$		
S_R	S_R	A_R	(1 1 0.64 0.04 0 0 0 0 0 0)
		A_P	(1 1 0.64 0.04 0 0 0 0 0 0)
		A_T	(0.4 0.4 0.4 0.04 0 0 0 0 0 0)
		A_Q	(0.8 0.8 0.64 0.04 0 0 0 0 0 0)
		A_S	(0.8 0.8 0.64 0.04 0 0 0 0 0 0)
	S_P	…	…

①大前提,②小前提,③结果向量.

(2) 在局部大前提为 $A_i \multimap\!\!\blacktriangleright S_j$ 时,抽取表中小前提为 A_k 所在的行,便得到一个矩阵 M_{ijk}(其中 i, j, k= 1,2,3,4,5)这样共得到 125 个矩阵,由这 125 个矩阵组成评价知识库,即集合$\{M_{111},\ldots,M_{ijk},\ldots,M_{555}\}$.它集中了在标准样本空间(通常为有限空间,且其各标准向量可划分为 5 个)中带有随机与模糊不确定因果状(变)态联系的全部信息,为完备有效地进行不确定自动推理和知识评价提供了可靠的依据.

3 认证逻辑的分析方法的应用

原理 **1(**一致性原理**).** 在客观世界中,在不确定性推理机制与大量样本统计下,因果关联规则在推理上的表征和在统计上的表征是一致的.

原理 **2(**适用性原理**).** 认证推理模式可适用于与因果关联规则相关的推理中.即

$$\frac{H \Rightarrow E \quad E}{H}$$

其中 H 为被检验的假设,可以视为经发掘后需要评价的因果关联规则 R.E 为从 H 可以推出的一些断言,可以视

为经检验得到的检验结果集合.在评价过程中所进行的检验是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,否则不满足一致性原理.这样检验结果的集合 E 中包含两个元素 E_1 和 E_2,E_1 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 $N(E_1)$;E_2 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 $N(E_2)$,并且满足 $N(E_1)+N(E_2)=N$.

根据正相关标准:E 认证 H,当且仅当 $Pr(H/E)>Pr(H)$.其中,$Pr(H)$为验前置信度,$Pr(H/E)$为验后置信度.这就是说,E 认证 H 当且仅当 H 相对于 E 的验后置信度大于其验前置信度.

将所发现的因果关联规则记为 $R(A_i \multimap S_j)$,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.这样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 $R(A_i \multimap S_j)$,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(可作为验前置信度).

验后置信度的确定取决于所进行的检验,在本评价过程中采用一致性原理作为检验的依据,使用的推理机制是因果关系定性推理机制.

定义 **7**. 将 $N(E_1)/(N(E_1)+N(E_2))$称为支持强度,记作 SUP.

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

4 评价算法(评价规则 $A_i \multimap S_j$)

在评价进行之前,首先要构建在原因语言场和结果语言场中的评价知识库,原因和结果的各个语言值对应的标准向量、各区间的中点和半径可以通过交互式,由领域专家和用户确定,对原有的根据情况进行修改.然后,按照第 2.2 节中的方法构建评价知识库,这个知识库可以作为所有原因为 A 的状(变)态和结果为 S 的状(变)态的规则的评价依据.然后取样本中原因 A 和结果 S 的数据,构成一个序偶的集合 $P=\{\langle t_w,s_w\rangle\}(w=1,2,\ldots,N)$,$t_w$ 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.设 $SUP_1=0$.在进行了上述处理之后,具体的评价步骤如下:

Step 1. 取原因的样本值 $t_w(w=1,2,\ldots,N)$,根据式(1)可得到因状(变)态输入向量 a_{tw}.

Step 2. 确定因状(变)态输入向量 a_{tw} 所属因状(变)态类型,如 $A_k(k=1,2,3,4,5)$.

Step 3. 以规则 $A_i \multimap S_j$ 作为局部大前提,以因状(变)态输入向量 a_{tw} 所属的因状(变)态标准向量 A_k 为小前提,可以在评价知识库中通过自组织的方式找到与其相匹配的唯一的知识矩阵 M_{ijk},根据自动推理模式(3)得到结果的状(变)态向量 S_{w1}.

Step 4. 计算 S_{w1} 所属的果状(变)态标准向量β.

Step 5. 对于序偶集 $P=\{\langle t_w,s_w\rangle\}$,取相应的结果的样本值 s_w,用模糊聚类的方法可得到它所属区间中的果状(变)态标准向量γ,如果$\beta=\gamma$,则 $SUP_1=SUP_1+1$,否则 $SUP_1=SUP_1$.

Step 6. 重复上述过程 N 次,得到 SUP_1.设 $SUP=SUP_1/N$,取规则的因果关联强度 CR 与其进行比较,若 $SUP>CR$,则规则被接受;若 $SUP \le CR$,则规则被拒绝.

5 实例运行检验

实例数据库是美国 1991 年某州社会调查结果中的部分数据.数据库内容包括调查对象的工作状况、婚姻状况、受教育年限、年收入状况等 17 个因素,数据库记录条数为 1500.我们针对前件为受教育年限,后件为年收入状况所发现的两条规则:“如果教育年限长 那么年收入多(记为 R_1)”和“如果教育年限长 那么年收入很多(记为 R_2)”,分别利用因果关联规则的评价方法进行评价.

在原因语言场中,语言变量为受教育年限,可划分为 5 个语言值:受教育年限很短(A_1)、受教育年限短(A_2)、受教育年限适中(A_3)、受教育年限长(A_4)、受教育年限很长(A_5).在论域中,最大值为 20(单位是年),最小值是 0.

各个语言值对应的标准样本点和半径可以由专家或用户来确定,我们将其分别取为 $A_1(1,2)$,$A_2(8.2,1)$,$A_3(11.8,1)$,$A_4(15,1)$,$A_5(17.8,0.8)$.然后要确定各个语言值对应的标准向量,因为我们使用的是Fuzzy语言场,那么只需确定受教育年限长(A_2)或短(A_4),其他的可以通过模糊变换得到.令 A_2=(1 0.8 0.6 0.4 0.2),A_4=(0.2 0.4 0.6 0.8 1),$A_1=(A_2)^2$=(1 0.64 0.36 0.16 0.04),$A_5=(A_4)^2$=(0 0.04 0.16 0.36 0.64),$A_3=(1-A_2)\wedge(1-A_4)$=(0 0.2 0.4 0.2 0).这些值可以通过根据数据的分布或经验得到.同理,在结果语言场中进行上述处理,得到年收入状况的5个语言值:年收入很少(S_1)、年收入少(S_2)、年收入中等(S_3)、年收入多(S_4)、年收入很多(S_5)的相应的标准样本点、半径和标准向量.根据第2.2中的方法构建评价知识库,上述两条规则的评价都使用这同一个评价知识库.

经过处理后,上述两条规则用定义中的形式分别表示为 R_1:[A_4]$\multimap$[S_4]和 R_2:[A_4]$\multimap$[S_5].$CR(R_1)$=0.205,$CR(R_2)$=0.264.然后分别进行评价,其结果为 $SUP(R_1)$=0.298>$CR(R_1)$,所以接受第1条规则;$SUP(R_2)$=0.106<$CR(R_2)$,所以不接受第2条规则为因果关联规则.评价的结果也比较符合人们的经验认识:受教育年限长是年收入长的一个原因,但是它并不是造成收入很长的直接原因.

6 与相关的工作的比较及进一步的工作

目前,对规则的主观感兴趣度的研究尚不多见.Piatetsky-Shapiro 和 Matheus[7]在健康保险领域研究了主观感兴趣的问题,并建立了知识发现系统 KEFIR.系统中的方法较好地解决了把主观感兴趣度与应用领域的结合问题,但是,这个方法不具有通用性,因为专家的领域知识以产生式规则的形式硬编码在系统中,所以这个系统不能用于其他领域.Silberschatz 和 Tuzhilin[8]提出使用信念和信念修正作为主观感兴趣的描述架.但是,他们只是给出了一些建议,没有实际的系统利用这个方法进行运行检验,Bing Liu[9]等人提出了使用用户期望的方法来找出用户真正感兴趣的规则,首先,要求用户根据他以往的知识或直觉给出期望的规则,然后使用模糊匹配技术将所发现的规则与期望的规则进行比较和匹配,根据匹配程度给它们排序,可根据需要决定输出符合用户期望程度高还是输出不符合用户期望高的规则.主要是针对于分类规则,并且要求输入的期望的形式与所发现的规则形式是一样的,即也是分类规则的形式,但在实际中,用户的知识并不都是一种类型的,这种方法实际上是根据与以往同类规则的相似程度来确定规则的感兴趣度.

本文所提出的方法首次将推理机制引入到规则的评价中,使评价逻辑性强,特别是对于因果关联规则来讲,最重要的是具有有效性,而其有效性主要体现在是否符合因果规律上.该评价方法较好地解决了因果规律这种重要的知识的表示(通过语言场的表示方法)和运用(通过采用因果关系定性推理机制)问题.我们今后的研究方向在于试图将场论的观点和方法应用到语言场理论中,找出各种规则形式在语言场和语言值的描述架下相应的推理形式,并将其推广到综合语言场中,进一步完善知识发现中的评价机制.

References:

[1] Agrawal, R., Mannila, H., Srikant, R., *et al*. Fast discovery of association rules. In: Fayyad, M., Piatetsky-Shapiro, G., Smyth, P., eds. Advances in Knowledge Discovery and Data Mining. Menlo Park, CA: AAAI/MIT Press, 1996. 307~328.

[2] Piatesket-Shapiro, G. Discovery, analysis, and presentation of strong rules. In: Piatesky-Shapiro, G., Frawley, W.J., eds. Advances in Knowledge Discovery and Data Mining. Menlo Park, CA: AAAI/MIT Press, 1991. 229~238.

[3] Symth, P., Goodman, R.M. An information theoretic approach to rule induction from databases. IEEE Transactions on Knowledge and Data Engineering, 1992,4(4):301~316.

[4] Toivonen, H., Klemettinen, M., Ronkainen, P., *et al*. Pruning and grouping discovered association rules. In: Mlnet Workshop on Statistics, Machine Learning, and Discovery in Database. 1995. 47~52. http://citeseer.nj.nec.com/toivonen95pruning.html.

[5] Yang Bing-ru. FIA and CASE based on fuzzy language field. Fuzzy Sets and Systems, 1998,95(2):83~89.

[6] Yang, Bing-ru. A valid mehod to judge FUZZY causal relation. In: The Selection of Mathematic Production. Tianjin: Tianjin Science and Technology Publishing House, 1983. 137~147 (in Chinese).

[7] Piatetsky-Shapiro, G., Matheus, C.J. The interestingness of deviations. In: Proceedings of the AAAI'94, Workshop on Knowledge Discovery in Databases. 1994. 25~36. http://citeseer.nj.nec.com/piatetsky-shapiro94interestingness.html.

[8] Silberschatz, A., Tuzhilin, A. What makes patterns interesting in knowledge discovery system. IEEE Transactions on Knowledge and Data Engineering, 1996,8(6):970~974.

[9] Liu, Bing, Hsu, Wynne, Mun, Lai-Fun, *et al*. Finding interesting patterns using user expectations. IEEE Transactions on Knowledge and Data Engineering, 1999,11(6):817~832.

附中文参考文献:

[6] 杨炳儒.FUZZY 因果联系的一种能行可判定方法.天津市数学研究成果选编.天津:天津科技出版社,1983.137~147.

An Evaluation Method for Causal Rules in KDD*

YANG Bing-ru, QI Yan-xia

(*School of Information Engineering*, *University of Science and Technology Beijing*, *Beijing* 100083, *China*)

E-mail: bryang@public.fhnet.cn.net

http:// www.ustb.edu.cn

Abstract: It is very important to evaluate the discovered rules in KDD (knowledge discovery in database). An evaluation method for causal rules is provided in this paper. The new and valid knowledge expression (language field and language value) and the reasoning mechanism (qualitative induction mechanism of causal relation) are used. The method is general and interactive. Its construction and the algorithm are given, and its validity is proved through case. By the comparison with the related work, it is proved to be an advanced method.

Key words: causal rules; measures of interestingness; evaluation; KDD (knowledge discovery in database)

* Received February 15, 2001; accepted June 5, 2001

Supported by the National Natural Science Foundation of China under Grant No.69835001; the National Science and Technology Key Foundation of the Ministry of Education of China under Grant No.[2000]175