doi:10.3969/j.issn.1673-4785.2009.03.007

评价信息元及其原信息元的获取方法

蔡 文,杨春燕
(广东工业大学 可拓工程研究所,广东 广州 510090)

摘 要:首先介绍评价对象与原对象、评价特征与原特征、评价量值与原量值的概念及其确定方法,进而给出评价信息元及综合关联度的构造,探讨从数据库中获取评价信息元的原信息元的方法,并给出相应的规则.该研究是基于可拓变换的可拓数据挖掘的基础.

关键词:评价信息元;原信息元;综合关联度;可拓数据挖掘
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-4785(2009)03-0234-05

A method for evaluation of information-elements and acquirement of the original information element

CAI Wen, YANG Chun-yan
(Research Institute of Extension Engineering, Guangdong University of Technology, Guangzhou 510090, China)

Abstract: This paper begins with an examination of the concepts of evaluated objects versus original objects; evaluated characteristics versus original characteristics; evaluated values versus original values. On this basis, a method to determine these parameters was developed. A construction method for evaluating information-elements and integrated dependent degrees was then proposed. Finally, methods for obtaining the original information-element of evaluated information-elements from a database were studied and the corresponding rules explored. This research provides a basic foundation for extension data mining based on extension transformation.

Keywords: evaluating information-element; original information-element; integrated dependent degree; extension data mining

数据挖掘是对数据库或数据仓库操作的方法[1-2].以关系数据库为例,其中的数据表可以与可拓学[3-4]中的基元集合[5-6]对应,把按基元形式组织的关系数据库中的三元数据(对象,特征或属性,量值)称为信息元.

人们处理问题时,对所涉及对象的评价需要使用某些特征,这些特征如何确定、如何获得相应的信息元、从数据库中找到确定这些信息元需要的特征和信息元,是必须研究的基本问题.而目前已有的数据挖掘还缺乏对这些基本问题的研究.因此,本文建立了评价所涉及的概念,讨论从数据库中获取评价信息元及其原信息元的规则和方法.这是基于可拓变换的可拓数据挖掘[7-8]的基础.

收稿日期:2008-04-25.
基金项目:国家自然科学基金资助项目(70671031);广东省普通高校人文社会科学研究重点资助项目(06ZD63008);广东省自然科学基金资助项目(05001832)
通信作者:蔡 文. Email: wcai@gdut.edu.cn.

1 基本概念

首先界定评价对象与原对象、评价特征与原特征、评价量值与原量值的概念,并介绍其确定方法.

1.1 评价对象与原对象

对具体的问题,必须确定需要评价的对象.评价对象可以是数据表中的对象,也可以是它们的组合或与表中对象相关的对象.评价对象确定以后,原始数据库中一切对它有影响的对象都称为该评价对象的原对象.例如,把上衣专柜作为评价对象,数据表的对象有各种款色的上衣,包括男上衣、女上衣、童装(上衣)等,它们叫做上衣专柜的原对象.

1.2 评价特征与原特征

一个事物有无数特征,在数据库中,储存着关于对象的很多特征及相应的量值.但在处理问题时,确定一个对象是否符合要求或是否具有某些性质的程度却不一定是已有的特征.例如,要修理天花板上的

日光灯时,需要的不是你的高度,而是“摸高”;考虑电梯需要多大功率的电机时,需要的数据不是所载货物和人的总重量,而是考虑了电梯“自重”后的曳引力;百货公司的商品中,电视机价格很高,日用品价格较低,因此,不能都用营业额作为营业员业绩的评价.

由此可见,在不同的目标下,评价某对象是否符合要求,是否具有某种性质,或者是否属于某种类型等,都有其特有的特征,这些特征称为评价特征. 数据库中储存的特征可以是评价特征,也可能不是评价特征;可能与评价特征有关,也可能与评价特征毫无关系. 不同的问题采用不同的评价特征,评价特征的确定者在不同的实际问题中也有所不同. 下面介绍若干种类型:

1)由问题的决策者确定评价特征.

一个新产品是否投入生产是由企业的决策者决定的,因此,他(或他们)会规定一些评价标准,如投资总额、生产周期和年利润值等. 公司录用职工时,人事部会规定一些评价指标,如年龄、性别、文化程度、学历等.

2)根据专业规范确定评价特征.

选择建筑材料时要依据建筑业的规范所规定的特征. 评价某人患何种疾病时要按照医学知识规定的评价特征. 评价各种事故时,也有专门规定的指标.

3)根据公用规范确定评价特征.

高考时确定能否录取的评价特征是总分数,对特殊专业有特殊的评价特征. 对一个地区环境的评价有联合国规定的评价特征.

4)利用常识确定评价特征.

出门穿衣的多少,一般使用“室外气温”作为评价特征. 走路抄近路,那是以“距离”或“所用时间”为评价特征;但也有的司机不是以“距离”作为挑选走哪条路的评价特征,而是以“路况”作为评价特征,在需要收费的路段,也有用“费用”作为评价特征的.

5)根据特殊爱好确定评价特征.

不同的人购买房子,有不同的评价标准,如有人用交通方便的程度,有人用环境噪音的程度,有人用楼层高低,有人用座向等等作为评价特征.

根据上述类型,可以对评价对象规定一定的评价特征,如百货公司把销售量和利润作为对销售业绩的评价特征. 确定了评价特征 $d_1,d_2,\cdots,d_q$ 以后,原始数据库中对这些特征有影响的特征称为该评价特征的原特征.

1.3　评价量值与原量值

如果评价对象 N 在数据表中的原始对象为 $O_i(i=1,2,\cdots,n)$,评价特征为 $d_l(l=1,2,\cdots,t)$,在数据表中的原特征为 $c_j(j=1,2,\cdots,m)$,评价对象 N 关于评价特征 d_l 的量值 u_l 由量值 $c_j(O_i)$ 确定,它们称为评价量值 u_l 的原量值. 如各类上衣的销售量就是上衣专柜的销售量的原量值.

1.4　评价信息元与原信息元

在确定了评价对象 O、评价特征 $d_l(l=1,2,\cdots,q)$ 和评价量值 $u_i(l=1,2,\cdots,q)$ 以后,信息元

$$I=\begin{bmatrix} N, & d_2, & u_1 \\ & d_2, & u_2 \\ & \vdots & \vdots \\ & d_q, & u_q \end{bmatrix} \triangleq (O,D,U)$$

称为评价信息元. 在数据库中的信息元集

$$\{I_i\}=\left\{ I_i=\begin{bmatrix} O_i, & c_1, & v_{i1} \\ & c_2, & v_{i2} \\ & \vdots & \vdots \\ & c_n, & v_{in} \end{bmatrix}, i=1,2,\cdots,m \right\}$$

称为评价信息元 I 的原信息元集或原信息元($i=1$ 时).

在数据库的变量中,有的没有评价特征对应的信息元. 因此,在可拓数据挖掘中,常常通过“原特征”来确定. 例如,要考虑哪些用户半年来没有使用电话,数据库中没有这样的信息元;但是可以用市话费与长途电话费之和(为0)或者市话通话时间与长途电话通话时间之和(为0)这样2个变量来获得. 市话费与长途电话费、市话通话时间与长途电话通话时间对应的信息元就是原信息元.

在可拓数据挖掘中,可以从评价特征和评价信息元出发,利用可拓学的方法,从数据库提供的原始信息元中,获取原信息元和原特征.

例如,对某销售服装的商场而言,若设评价对象为“上衣专柜”,评价特征为“利润值”和“销售量”,则评价信息元为

$$I=\begin{bmatrix} \text{上衣专柜}, & \text{利润值}, & u_1 \\ & \text{销售量}, & u_2 \end{bmatrix}$$

原信息元集为 $\{I_i\}=\{I_1,I_2,I_3,\cdots\}$,其中:

$$I_1=\begin{bmatrix} \text{长袖上衣}, & \text{价格}, & v_{11} \\ & \text{销售量}, & v_{12} \end{bmatrix},$$

$$I_2=\begin{bmatrix} \text{短袖上衣}, & \text{价格}, & v_{21} \\ & \text{销售量}, & v_{22} \end{bmatrix},$$

$$I_3=\begin{bmatrix} \text{羊毛上衣}, & \text{价格}, & v_{31} \\ & \text{销售量}, & v_{32} \end{bmatrix}.$$

$$\vdots$$

2 综合关联度

评价特征有时只有一个,有时有多个,对于多个评价特征,必须进行综合评价.

设关于某问题的评价指标有多个,为了综合评价一个信息元 I_i 符合要求的程度,建立了综合关联度的概念.

根据决策者或者专业知识,规定评价特征为 d_1, $d_2,\cdots,d_q$,根据这些特征计算出评价特征的量值.

给定信息元集

$$\{I_i\} = \{I_i \mid I_i = (O_i,C,V_i), i = 1,2,\cdots,n\},$$

若 $d_1,d_2,\cdots,d_q$ 为评价特征,得到 I_i 的评价信息元

$$D_i = \begin{bmatrix} I_i, & d_1, & u_{i1} \\ & d_2, & u_{i2} \\ & \vdots & \vdots \\ & d_q, & u_{iq} \end{bmatrix}.$$

在评价特征 $d_p(p=1,2,\cdots,q)$ 的量域 $V(d_p)$ 上建立关联函数 $k_p(x)$,对 $D_{ip}=(I_1,d_p,u_{ip})$ 计算 $k_p(D_{ip}) = k_p(u_{ip})$,规范化得到关联数

$$k_{ip} = \frac{k_p(D_{ip})}{\max\limits_{p\in\{1,2,\cdots,q\}\mid k_p(D_{ip})\mid}},$$

$$p = 1,2,\cdots,q, i = 1,2,\cdots,n.$$

称 $K(I_i)$ 为关于 I_i 的综合关联度. 根据实际问题,$K(I_i)$ 可以取如下3种类型之一[9]:

1) 若只有 I_i 关于每个评价特征的值都符合要求,才算 I_i 符合要求,则取

$$K(I_i) = \bigwedge_{p=1}^{q} k_{ip},$$

它表示 $K(I_i)$ 取 $k_{i1},k_{i2},\cdots,k_{iq}$ 的最小值.

2) 若只要 I_i 关于某一个评价特征的值符合要求,I_i 就符合要求,则取

$$K(I_i) = \bigvee_{p=1}^{q} k_{ip},$$

它表示 $K(I_i)$ 取 $k_{i1},k_{i2},\cdots,k_{iq}$ 的最大值.

3) 如果实际问题对评价特征侧重的程度有所不同,以权系数 α_p ($\sum\limits_{p=1}^{q}\alpha_p = 1$) 表示该特征的重要程度,即取

$$K(I_i) = \sum_{p=1}^{q} \alpha_p k_{ip}.$$

当评价特征只有1个时,即

$$D_i = (I_i,d,u).$$

建立 $V(d)$ 上的关联函数 $k(x)$,若规范关联数为 $k(u)$,规定综合关联度为

$$K(I_i) = k(u).$$

在多评价特征信息元可拓集中,对论域 W 中的每一信息元 I_i,若 $K(I_i)>0$,则认为信息元 I_i 符合要求;若 $K(I_i)<0$,则认为信息元 I_i 不符合要求;若 $K(I_i)=0$,则具体问题具体分析,因为有些实际问题需要把零界元素作为符合要求的信息元,而另一些实际问题则不然.

若存在信息元 I_0,满足

$$K(I_0) = \max_{1\leqslant i\leqslant n}\{K(I_i), I_i \in S\},$$

则表示 S 中 I_0 的综合关联度最大,即优度(符合综合要求的程度)最高,可为处理矛盾问题提供定量的依据.

3 从数据库中获取评价信息元的原信息元的方法

由于问题的评价特征和数据库中已有的特征有区别,也有联系,因此必须讨论如何从数据库中获取评价信息元所需要的特征和量值.

3.1 利用基元的拓展性确定原信息元

在可拓数据挖掘中,通过评价信息元来挖掘分类知识、变换的传导知识. 但从已有数据库中可能找到与评价信息元相同的信息元,也可能找不到. 因此,需要讨论获取评价信息元的原信息元的方法.

把变量用信息元规范表示以后,可以利用基元的拓展性去确定评价信息元的原信息元.

3.1.1 用信息元的相关性确定原信息元

事物的一个特征,往往有其相关的特征,例如,若评价特征为电费 d,而某企业的电费与用电量 c 是相关的. 如果数据库中没有电费 d 这一特征,但有该公司用电量的数据,那么可以用 c 对应的量值来计算 d 对应的量值,c 就是 d 的原特征.

通常数据库中有与评价特征相关的特征. 因此,可以根据信息元的相关性,获得与评价信息元相关的原信息元. 根据基元的相关性,有如下规则:

规则1 给定信息元集

$$\{I_i\} = \{I_i \mid I_i = (O_i,C,V_i), i = 1,2,\cdots,n\}$$

其中:

$$C = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{bmatrix}, V_i = \begin{bmatrix} V_{i1} \\ V_{i2} \\ \vdots \\ V_{im} \end{bmatrix}.$$

若评价特征 d 的原特征为 c' 或 $c'_1,c'_2,\cdots,c'_r$,评价信息元 $D=(I,d,d(I))$ 满足

$$d(I) = f(c'(O))$$

或

$$d(I) = f(c'_1(O),c'_2(O),\cdots,c'_r(O)),$$

则 $I=(O,c',c'(O))$ 或

$$I=\begin{bmatrix} O, & c_1', & c_1'(O) \\ & c_2', & c_2'(O) \\ & \vdots & \vdots \\ & c_r', & c_r'(O) \end{bmatrix}$$

是评价信息元 $D=(I,d,d(I))$ 的原信息元.

若评价特征 d 的相关特征为 d',d' 又与 C 中的特征 c' 相关,由相关性的传递性,d 与 c' 相关,因此,c' 可以是 d 的原特征. 进而,若 d 的相关特征集为 $\{d'\}$,数据库中的特征集为 $\{c\}$,则 $\{d'\}\cap\{c\}=\{c(d')\}$ 中的特征也是 d 的原特征. 由此可见,要找到数据库中 d 的原特征,可以先求出 d 的相关特征集,再找它与 $\{c\}$ 的交集.

3.1.2　用信息元的可扩性确定原信息元

利用信息元的可结合性和可分解性,可以从数据库得到评价信息元的原信息元. 例如,对移动电话的通讯数据而言,可以按以下几类消费者进行分类:

1)稳定型消费者,他们使用移动电话不随时间的变化而变化;

2)增长型消费者,他们使用移动电话的时间不断增加;

3)接受型消费者,他们大多只接收电话;

4)发送型消费者,他们大多数情况下向外打电话.

那么,如何确定这4类消费者呢? 可以用"使用电话稳定度"和"使用时间增长度"这2个评价特征. 但数据库中并没有这2个特征,却有"通话次数"和"通话时间"2个特征,利用后两个特征的信息元按月份进行分解,再汇总,可以得到所需要的2个评价特征的原信息元.

与相关特征类似,对于评价特征 d,可以先求出 d 的可扩特征集 $\{d_r\}$,那么,$\{c(d_r)\}=\{d_r\}\cap\{c\}$ 中的特征可以作为 d 的原特征,再计算出相应的原信息元和评价特征元.

以下规则中的符号,如无特别说明,均与规则1相同,此处不再赘述.

根据信息元的可扩性,有如下规则:

规则2　给定信息元集

$\{I_i\}=\{I_i\mid I_i=(O_i,C,V_i),i=1,2,\cdots,n\}$,

若 $c_{j_1}\cdot c_{j_2}=d$,即 c_{j_1} 和 c_{j_2} 是 d 的原特征,则 (O_i,c_{j_1},v_{ij_1}) 和 (O_i,C_{j_2},v_{ij_2}) 是 (O_i,d,u) 的原信息元,且

$$(I_i,d,u)=(O_i,c_{j_1}\cdot c_{j_2},v_{ij_1}\cdot v_{ij_2}).$$

规则3　给定信息元集

$\{I_i\}=\{I_i\mid I_i=(O_i,C,V_i),i=,1,2,\cdots,n\}$,

c_j 是 d 的原特征,且 $O_1\otimes O_2\otimes\cdots\otimes O_r=O$,则 $(O,c_j,v_1\otimes v_2\otimes\cdots\otimes v_r)$ 是 (I,d,u) 的原信息元,且

$$(I,d,u)=(O_1\otimes O_2\otimes\cdots\otimes O_r,c_j,v_1\otimes v_2\otimes\cdots\otimes v_n).$$

例如,某企业的电费这一评价特征对应的评价信息元可以用各部门用电量之和与电价之积来计算. 因而,各部门和它们用电量组成的信息元就是该企业和电费组成的信息元的原信息元.

3.1.3　用信息元的蕴含性确定原信息元

有的评价信息元在已有的数据库中无法得到,但利用信息元的蕴含性,可以从数据库中找到相应的原信息元. 根据信息元的蕴含性,有如下规则:

规则4　给定信息元集

$\{I_i\}=\{I_i\mid I_i=(O_i,C,V_i),i=1,2,\cdots,n\}$,

若

$$D=(O,d,u)\Leftarrow(O_i',c',v_i')=I_i',$$

则 I_i' 是 D 的原信息元,c' 是 d 的原特征.

3.1.4　用信息元的发散性确定原信息元

利用信息元的发散性,也可以获得评价信息元的原信息元,有如下规则:

规则5　给定信息元集

$\{I_i\}=\{I_i\mid I_i=(O_i,C,V_i),i=1,2,\cdots,n\}$,

和评价信息元

$$D=(I,d,u)=[(O,c,v),d,u],$$

若

$$I=(O,c,v)\text{—}\mid I'=(O_{i_0},c,v)\text{—}\mid I''=(O_{i_0},C,V_{i_0})\in\{I_i\},$$

其中 $i_0\in\{i=1,2,\cdots,n\}$,则 I'' 是 D 的原信息元.

3.2　**寻找原信息元的注意事项**

寻找原信息元要依靠人们和计算机的知识库中的知识,包括:常识、公式、领域知识等. 在可拓数据挖掘工作中,不断积累这些原特征和原信息元,可以为下一次挖掘工作服务.

寻找原信息元是为了挖掘有用的知识. 因此,对所得到的原信息元必须进行评价,选择优者,淘汰劣者. 这种评价可以利用已有的知识、历史资料和数据挖掘得到的专业知识以及优度评价法综合处理. 有时,若干原信息元可以表示同一件事情,则择其优者而用之.

4　结束语

本文研究了评价信息元及其相应的原信息元的获取方法与规则,这是可拓数据挖掘的基础工作,为从数据库中获取可拓分类知识提供了理论基础.

参考文献:

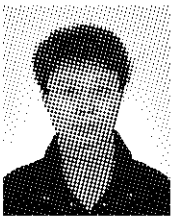
[1]陈　安,陈　宁,周龙骤,等. 数据挖掘技术及应用[M]. 北京:科学出版社,2006.

[2]陈文伟. 数据仓库与数据挖掘教程[M]. 北京:清华大

学出版社,2006.
[3]CAI Wen. Extension theory and its application[J]. Chinese Science Bulletin,1999,44(17):1538-1548.
[4]蔡 文,杨春燕,何 斌. 可拓逻辑初步[M]. 北京:科学出版社,2003.
[5]杨春燕,蔡 文. 可拓工程[M]. 北京:科学出版社,2007.
[6]李立希,杨春燕,李铧汶. 可拓策略生成系统[M]. 北京:科学出版社,2006.
[7]蔡 文. 变化的知识与可拓数据挖掘[C]//中国人工智能进展(2007). 北京:北京邮电大学出版社,2007,12:951-955.
CAI Wen. Study on changing knowledge and extension data mining[C]//Proceedings of 2007 National Conference on Artificial Intelligence. Beijing: Beijing University of Posts and Telecommunications(BUPT) Publishing House, 2007, 12: 951-955.
[8]YANG Chunyan. Extension classification method and its application based on extensible set[C]//Proceedings of 2007 International Conference on Wavelet Analysis and Pattern Recognition. Beijing, 2007, 11: 819-824.
[9]杨春燕. 多评价特征基元可拓集研究[J]. 数学的实践与认识,2005,35(9):203-208.
YANG Chunyan. Study on the basic-element extension set of multi evaluating characteristics[J]. Mathematics in Practice and Theory, 2005, 35(9): 203-208.

作者简介:

蔡 文,男,1942年生,研究员,国家级有突出贡献的专家,可拓学的创立者,中国人工智能学会常务理事,中国人工智能学会可拓工程专业委员会主任. 主要研究方向为可拓学、人工智能、决策科学,主持国家基金项目5项,参加多项国家基金项目. 发表学术论文多篇,出版专著7部.



杨春燕,女,1964年生,研究员,中国人工智能学会理事,中国人工智能学会可拓工程专业委员会常务副主任. 主要研究方向为可拓学、人工智能、决策科学,主持国家基金项目2项,广东省自然科学基金项目2项,参加多项国家基金项目. 发表学术论文40余篇,出版专著7部.

The Tenth IASTED International Conference on Artificial Intelligence and Applications

第10届IASTED人工智能及应用国际学术会议

February 15–17, 2010

Innsbruck, Austria

The Tenth IASTED International Conference on Artificial Intelligence and Applications (AIA 2010) will focus on bringing together researchers and practitioners working in the field of theory and practice of computer learning, pattern recognition, inductive and transductive inference, graphical models, neural networks, and neuroscience. These topics are now central in Artificial Intelligence, and communication between different groups of people involved in natural and computer learning has become crucial for further progress in AI.

The proceedings will be sent for indexing in Cambridge Scientific Abstracts, Conference Proceedings Citation Index, EI Compendex, FIZ Karlsruhe, INSPEC.

Important deadlines:

Submissions due: September 15, 2009

Notification of acceptance: November 1, 2009

Final manuscripts due: November 15, 2009

Registration deadline: December 1, 2009

Contact us

For more information, or to be placed on our mailing list, please contact:

IASTED Secretariat – AIA 2010

B6, Suite 101, Dieppe Avenue SW

Calgary, AB, Canada T3E 7J9

Tel: 403-288-1195 Fax: 403-247-6851

E-mail: calgary@iasted.org Website: www.iasted.org