数据挖掘及其在通信企业经营分析领域的应用

刘 蓉

(长沙理工大学,湖南 长沙 410007)

摘 要:从数据挖掘(DM)的分析模型出发,提出了一种新的DM模型,包括多个多维立方体模型和一系列多维算子模型,并将此DM模型应用于通信企业消费者行为分析中,提供了消费分析的实例和方法。

关键词:数据挖掘;多维立方体数据模型;多维算子;通信企业消费者行为分析

中图分类号:F606 **文献标识码**:A **文章编号**:1001-7348(2004)10-0159-03

数据挖掘(DM)是基于数据仓库(DW)的一种多维分析技术,它能为企业经营决策提供技术支撑。数据挖掘通过对海量信息的筛选和处理,借用先进管理决策模型,可以大大提高企业经营决策的准确性、实时性和事实性。

Anindya Datta 等在文献中提出了一种多维立方体数据模型以及用以支持在此数据模型上实现DM分析的代数模型,在研究利用DM进行通信企业经营分析的过程中,我们通过增加关联算子在跨多个多维立方体的分析方面扩展了这个代数模型,例如跨多个多维立方体(或其切片)的聚集运算和关联运算。

1 DM分析模型

1.1 多维立方体数据模型

一个多维立方体是一种逻辑上的数据组织方式,是实现DM的数据对象。多维立方体在这个DM模型中作为下面所提的多维立方体算子模型的基本输入和输出。一个多维立方体可以定义为一个四元组(D,M,A,f),这4个元素从不同角度描述了多维立方体的特性,它们的定义分别如下:

$D=(d_1,d_2,\cdots,d_n)$是维集,其中d_i是维的名称,来自域$dom_{dim(i)}$;

$M=(m,m_2,\cdots,m_k)$是测度集,其中m_i是测度的名称,来自域$dom_{measure}(i)$;

$A=(a_1,a_2,\cdots,a_t)$是属性值,其中a_i是属性的名称,来自域$dom_{attr}(i)$;

$f:D\rightarrow A$是维集到属性集的一对多映射,也就是说,对应于每个维有一组属性,约束条件如下:

(1)$DIM=0$,即维集和测度集是没有交集的;(2)$\forall i,j,i\neq j,f(d_i)If(d_j)=0$,即任意两个维的属性集是没有交集的。

上述关于多维立方体的定义是一个抽象定义,具体针对一个多维立方体实例,可以用一个六元组(D,M,A,f,v,g)来定义。一个多维立方体实例是将一个抽象的多维立方体进行实例化得到的,其中前4个元素前面已经解释过了。那么,v是一个值的集合,任意一个元素$v_i\in v$都可以用一个K元组$(\mu_1,\mu_2,\cdots,\mu_k)$来表示,其中$\mu_i$是第$i$个测度$m_i$的实例化;$g$代表映射$g:dom_{dim(1)}\times dom_{dim(2)}\times\cdots\times dom_{dim(n)}\rightarrow v$,即$g$映射通过将多维立方体中的单元与值集$v$中的元素进行关联,实现这些单元的实例化。

1.2 多维算子模型

多维算子模型包括一系列定义在多维立方体上的算子,它们以多维立方体作为基本输入和输出,实现DM的基本代数。下面介绍这些主要的算子模型。

(1)限定算子(σ)。

限定算子在一个或多个维上按照一定的维条件对多维立方体进行限定。假定p是一个元谓词,表示一个维上限定条件的逻辑表达式,而复合谓词$p=p_1(op)p_2(op)\cdots p_i$是一系列元谓词的逻辑表达式,其中$op$是逻辑算子。

输入:多维立方体$C_1=(D,M,A,f,v,g)$和表示限定条件的复合谓词p;

输出:多维立方体$C_0=(D_0,M_0,A_0,f_0,v_0,g_0)$,其中$D_0=D,M_0=M,A_0=A,f=f,v_0\subseteq v,g_0$使得每一个$g_0^{-1}(v_0)$的元素满足$P$;代数表达:$\sigma_p(C_1)=C_0$可以看出限定算子是支持DM分析的切片和切块分析运算的基本算子。

(2)聚集算子(σ)。

聚集算子在一个或者多个维上进行聚集运算。假定h是定义在某个测度m_i上的聚集函数。S是一个聚集目标属性集$\{a_1,a_2,\cdots,a_q\}$,且有$S\subseteq A$,聚集运算将在A中除了这些属性的其他属性上进行。映射$\delta:A\rightarrow D$表示一个将上述S中的每个属性a_i与其相关的维d_i关联起来的一一映射。下面给出聚集算子的代数描述:

输入:多维立方体$C_1=(D,M,A,f,v,g)$和一个用于聚集的测度m_i,以及一个聚集目

收稿日期:2004-04-08

标属性集 S；

输出：多维立方体 $C_0=(D_0,M_0,A_0,f_0,v_0,g_0)$，其中 $D_0=\{d_1,d_2,\Lambda,d_q\}$，$q=|S|$ 且 $\forall a_i\in S$，$d_i=\delta(a_i)$；$M_0=\{m_i\}$；$A_0=Y\ \forall_{d_i}\in D_0 f(d_i)$；$f_0=f$；$v_0$ 是在 v 的元素上进行聚集运算后得到的值集；$g_0:dom_{\dim(1)}\times dom_{\dim(z)}\times\Lambda\times dom_{\dim(q)}\rightarrow v_0$。代数表达：$a_{k,m_i,s}(C_1)=C_0$。

(3)关联算子(θ)。

在文献[2]中未提及关联算子。我们提出的关联算子是一种跨多维立方体的聚集运算，由二个或以上的多维立方体在某个维点上的 K 个切片运算得到另一个新的多维立方体某维上的多个切片，新多维立方体的其它维可以沿用旧多维立方体的维来构成。输出的多维立方体为 C_0，输入的多维立方体为 $C_{1_1},C_{1_2},C_{1_3},\cdots,C_{1n}$。对于 C_0，除了做切片的维(设为第 j 维，记为 d_0)与 C_{1_1} 不同之外，其余的维与 C_{1_1} 相同，主要是用于实例化值集的多边关联。关联算子和聚集算子有某种相似性，如关联算子是针对测度在某个、多个多维立方体 C，或在 C 的内部某个(或某些)维上进行聚集；而聚集算子只进行内部聚集。S 是一个聚集目标属性集$\{a_1,a_2,\cdots,a_n\}$，因此有 $S\subseteq A$，因此聚集算子的结果满足 $D_0\subseteq D,M_0\subseteq M,A_0\subseteq A,f_0=f,v_0\subseteq v$；而关联算子是外部聚集，是由多个多维立方体生成另一个新的多维立方体。下面给出关联算子的代数描述：假定 l 是定义在某个测度 m_i 上的关联函数，元谓词 P_{jk_n} 表示在第 n 个多维立方体 C_{1n} 的某个维 d_{j_n} 上获得切片 k_n 的限定条件逻辑表达式，(其中 $n\in(1,2,\cdots,n)$表示共有 n 个多维立方体)。l 是在维 $d_{j_1},d_{j_2},\cdots d_{j_n}$ 等多个维上根据上述限定条件获得的切片进行关联运算的函数。元谓词 q_{j_1} 表示在输出多维立方体 C_0 上的维 d_{0_j} 上获得切片 t 的限定逻辑表达式。

输入：多个多维立方体 $C_{1_1},C_{1_2},C_{1_3},\cdots,C_{1_n}$，其中 $C_{1_1},C_{1_2},C_{1_3},\cdots,C_{1n}=(D,M,A,f,v,g)$；一个用于聚集的测度 m_i；进行关联运算的维 $d_{j_1},d_{j_2},\cdots,d_{j_n}$；以及元谓词 p_{j_n} 和元谓词 q_{j_1}。

输出：多维立方体 $C_0=(D_0,M_0,A_0,f_0,v_0,g_0)$ 的切片 t，即 $\theta_{q_{j_1}}(C_0)$，其中 D_0 是将 D_{1_1} 中 d_{j_1} 替换为 d_{0_j} 而得到，A_0 是将与 d_{j_1} 相关的维属性替换为与 d_{0_j} 相关的维属性得到的，v_0 是在 v 的元素上进行关联运算后得到的值集。

$g_0:dom_{\dim(1)}\times\Lambda\times dom_{\dim(j-1)}\times dom_{\dim(j)}\times dom_{\dim(0_j)}\times dom_{\dim(j+1)}\times\Lambda\times dom_{\dim(n)}\rightarrow v_0$；代数表达式：

$\theta_{1,m_i,d_j}(\sigma_{p_{j_1}}(C_{1_1},C_{1_2},C_{1_3},\cdots,C_{1n}),\sigma_{p_{j_2}}(C_{1_1},C_{1_2},C_{1_3},\cdots,C_{1n})\Lambda\sigma_{p_{j_n}}(C_{1_1},C_{1_2},C_{1_3},\cdots,C_{1n})=\sigma_{q_{j_1}}(C_0)$回溯运算是关联运算的逆运算，是根据运算结果切片和由关联函数确定的维限定条件，获取参与运算的所有多维立方体切片，记作：$\theta^{-1}_{1,m_i,d_j}(\sigma_{q_{j1}}(C_{1_1},C_{1_2},C_{1_3},\cdots,C_{1n}))=(\sigma_{p_{j1}}(C_0),\sigma_{p_{j2}}(C_0),\Lambda,\sigma_{p_{jn}}(C_0))$

(4)分隔算子(γ)。

分隔运算是将多维数据立方体按照一定的标准分隔成为有意义的组。分隔运算在特定的聚集运算和关联运算中都是必需的。定义映射 $t:dom_{\dim(1)}\times dom_{\dim(2)}\times\Lambda\times dom_{\dim(n)}\rightarrow V_G$($V_G$ 是一个值组的集合)。定义 R 是一个分隔维属性的集合，有 $R\subseteq A$。分隔算子的代数描述如下：

输入：多维立方体 $C_1=(D,M,A,f,v,g)$，一个分隔维属性集合 R 和一个分隔函数 t；

输出：多维立方体 $C_0=(D_0,M_0,A_0,f_0,v_0,g_0)$，其中 $D_0=D,M_0=M,\forall a_i\in R,A_0=AYt(a_i)$，$f_0=f,|v_0|=|v|$；

代数表达式：$\gamma_{t,R}(C_1)=C_0$

2 以电信企业的“客户消费行为”分析为例，讨论电信企业经营分析的数据挖掘实现

2.1 电信企业客户消费分析的多维立方体数据模型

电信企业经营分析依赖于大量基础数据，以“客户消费行为”分析为例，客户消费和欠费数据是电信基础数据的一部分。它们来自电信九七工程(营业系统)和计费系统等业务运行数据库中，反映了某类客户对某种电信产品在某一时间范围内的消费值和欠费值。而消费比较数据比消费和欠费等基础数据要高一层，称为准基础数据(作为示例，本文我们仅考虑以消费比较数据作为准基础数据)，它是通过基础数据得到的。它反映了客户的相互关系和内在联系，反映某类或某个客户的消费水平、消费潜力、消费倾向、消费特征和消费关联等情况。对于电信企业来说，已经有一整套关于客户消费行为分析的基础和准基础数据的指标体系(可查阅有关电信企业经营分析指标)。此节，本文应用上文提出的 DM 分析模型及其代数算子，来获取电信客户消费行为分析的有关指标。下面，我们分别对电信客户消费分析的基础数据(消费数据和欠费数据)和准基础数据(客户消费比较)进行数据建模，分别用 Consumel_level (客户消费水平)、Owe_level (客户欠费水平)和 consume_comp(客户消费比较)表示这三个多维立方体。定义如下：

(1)$Consume\_level=(D_1,M_1,A_1,f_1,v_1,g_1)$，其中

$D_1=(Time,Name,Consume\_index)$，电信消费水平通常是从时间、客户名称(表示某类客户或某个客户)和客户所消费的产品等 3 维进行；

$M_1=(Actual,Forecast)$，人们关心的消费水平数据通常有两类，一类是实际值，一类是预测值，可以看出 $D_1|M_1=0$；

$A_1=(Day,Month,Year,Group\_name,Singal\_name,Index\_table,Index\_name)$；

$f_1(Time)=\{Day,Month,Year\}$，时间维可以用日、月、年属性来描述；

$f_1(Name)=\{Group\_name,Singal\_name\}$，客户名称维可以用集团客户或单个客户的名称 2 个属性来描述；

$f_1(Consume\_index)=\{Index\_table,Index\_name\}$，消费产品名称维可以用产品表和产品名称 2 个属性来描述；

$\forall i,j,i\neq j,f_1(d_i)|f_1(d_j)=0$；

v_1 是一个二元组的值集，v_1 对上述 4 个元素确定的多维立方体进行实例化，每一个元素是 $(Actual,Forecast)$ 的一个实例；g_1 表示映射 $g_1:dom_{\dim(Time)}\times dom_{\dim(Name)}\times dom_{\dim(Consume\_index)}\rightarrow v_1$

(2)$Owe\_level=(D_2,M_2,A_2,f_2,v_2,g_2)$，其中

$D_2=(Time,Name,Owe\_index)$，客户欠费水平通常从时间、客户名称和欠费的电信产品等 3 个维来描述；

$M_2=(Actual,Forecast)$，人们关心的客户欠费水平数据通常有两类，一类是实际值，一类是预测值，可以看出其满足 $D_1|M_1=0$；

$A_2=(Day,Month,Year,Group\_name,Singal\_name,Index\_table,Index\_name)$；

$f_2(Time)=\{Day,Month,Year\}$，时间维可以用日、月、年属性来描述；

$f_2(Name)=\{Group\_name,Singal\_name\}$，客户名称维可以用集团客户或单个客户的名称来描述；

$f_2(Owe\_index)=\{Index\_table,Inddex\text{-}name$

}，欠费的电信产品由欠费表和欠费产品名等2个属性来描述；

v_2 是一元组的值集，v_2 对上述4个元素所确定的多维立方体进行实例化，每一个元素是 *Owe-level* 的(*Actual*,*Forecast*)的一个实例；g_2 表示映射 $g_2:dom_{dim\ [Time]}\times dom_{dim\ [Name]}\times dom_{dim(Owe\_index)}\rightarrow v_2$。

(3)$consume\_comp=(D_3,M_3,A_3,f_3,v_3,g_3)$

D_3=(*Time*,*Name*,*Comp_index*)，客户消费比较可以从时间、客户姓名和消费比较指标等3个属性来描述；

M_3=(*Comp_value*),*Comp_value* 表示客户消费比较指标的值；

A_3 =(*Day*,*Month*,*Year*,*Group_name*,*Singal_name*,*Analyze_kind*,*Comp_index*);

f_3=(*Time*)={*Day*,*Month*,*Year*},时间维可以用日、月、年属性来描述；

f_3(*Name*)={*Group_name*,*Singal_name*}，名称维可用集团客户和单个客户的名称等2个属性描述；

f_3 (*Comp_index*) ={*Analyze_kind*,*Index_name*}，消费比较指标可以用消费分析方法和消费比较指标名称等2个属性来描述；同样满足 $\forall i,j,i\neq j,f_3(di)\cap f_3(d_i)=0$；

v_3 是个一元组的值集，v_3 对上述4个元素确定的多维立方体进行实例比，每一个元素是(*Comp-value*)的一个实例；

g_3 表示映射 $g_3:dom_{dim\ (Time)}\times dom_{dim\ (Name)}\times dom_{dim(Comp\_index)}\rightarrow v_3$

2.2 数据挖掘算子在电信客户消费分析中的应用

(1)横向比较分析。横向比较分析就是根据电信客户消费的基本数据，通过计算一些比较值来获得对客户消费特征的把握。比较分析通常是计算客户消费的两个产品之间消费值的绝对值或平均值之间的相对比率。比较分析就是由多维立方体 *Consume_level* 和 *Owe_level* 生成另一个多维立方体 *Consume_comp* 的情况)。

客户对电话和上网两种产品的消费值比较(简称电话/上网比较值)=电话消费/上网消费。客户对电话和上网两种产品的消费值分别是多维立方体 *Consume_level* 的维 *Consume_index* 上属性 *Index_name* 的两个实例值。两种产品消费比较值是多维立方体 *Consume_comp* 的维 *Consume_index* 上属性 *Index_name* 的一个实例值。可以用关联算子和限定算子实现该比较分析。

(2)纵向比较分析。对电信企业来说，还可以从纵向角度进行消费的比较分析，如对客户消费进行历史趋势分析。用聚集算子和限定算子便可实现以上的对比分析，下面给一个例子：

"给出1993~2003年10年间某电信企业客户关于电话产品的消费变化趋势"，这是从时间维上进行比较，在客户名称维上进行聚集运算。代数表达式如下：$\sigma_{(year\geq 1993\wedge year\leq 2003\wedge Index\_name="总销售值")}(\alpha_{sum(Actual),Year}(C_1))=C_0$。

(3)回溯分析。回溯分析是一种分解结果的方法。通过对影响结果的各个因素的分解，明确指标变动的原因是由哪个因素造成的，从而为采取措施指明方向。请看示例：

长话占总话费消费的比重(简称长话比重)=长话费/总话费，显然长话比重是通过长话消费和总话费消费数据关联得到的，存在如下关联运算：

$\theta_{DIV,Actual,Consume\_index}(\sigma_{Consume\_index1}(Consume\_level),\sigma_{qComp\_index_j}(Consume\_level))=\sigma_{qComp\_index_j}(Consume\_comp)$

其中，$PConsume\_index_1$=(*Index_name*="长话消费")；$PConsume\_index_2$=(*Index_name*="总话费")

$qComp\_index_j$=(*Index_name*="长话比重")

$\theta^{-1}DIV,Actual,Consume\_index\ (\sigma_{qComp\_index_j}(Consume\_comp))=(\sigma_{pConsume\_index_1}(Consume\_level),\sigma_{PConsume\_index_2}(Consume\_level))$。

上述指标可用回溯运算来分解。

(4)较复杂的比较分析。我们所指的较复杂的比较分析，是指由输入的多个多维立方体的某些维进行比较关联而得到的一个新的输出多维立方体的比较分析过程。示例如下：

在电信企业的客户消费分析中，虚假消费率=欠费/总消费，通过虚假消费率的分析，电信企业可以识别某类或某个有欺诈行为和倾向的客户，以便采取及时行为，留住有利润的客户，转化有摩擦的客户，淘汰虚假客户。显然，虚假消费率是通过欠费和总消费这两上值关联得到的。欠费来自多维立方体 *Owe_level* 的维 *Owe_index* 的属性 *Index_name* 的一个关于 *Actual* 的实值，总消费来自多维立方体 *Consume_level* 的维 *Consume_index* 的属性 *Index_name* 的一个关于 *Actual* 的实值，总消费来自多维立方体 *Consume_level* 的维 *Consume_index* 的属性 *Index_name* 的一个关于 *Actual* 的实值，总消费来自多维立方体 *Consume_comp* 的维 *Comp_index* 的属性 *Index_name* 的一个关于 *Actual* 的实值。虚假消费率是多维立方体 *Consume_comp* 的维 *Comp_index* 的属性 *Index_name* 的一个 *Comp_value* 实值。

$\theta_{DIV,Actual}\ (\sigma_{PConsume\_index}(Consume\_level),\sigma_{POwe\_index}(Owe\_level))=\sigma_{q,Consume\_Comp_j}(Consume\_comp)$

其中，$pConsume\_index$ =(*Index_name* ="总消费")；

$POwe\_index$=(*Index_name*="欠费")

$q_{Comp\_index_j}$=(*Index_name*="虚假消费率")

上述运算的回溯运算可以由下式实现

$\theta^{-1}DIV,Actual\ (\sigma_{qComp\_index_j}(Consume\_comp))=(\sigma_{PConsume\_index}(Consume\_level),\sigma_{POwe\_index}(Owe\_level))$

3 结束语

综上所述，本文在文献[2]的基础上，提出了一种基于多维立方体数据模型和多维算子模型的数据挖掘方法，在跨多个多维立方体计算方面提出了关联算子模型，并将此模型应用于通信企业的客户消费行为分析中，为通信企业经营分析提供了实例化的研究方法。在今后工作中，我们将进一步完善和丰富本文所提出的这一数据挖掘模型，探讨在通信企业经营分析的其它领域使用这一分析决策模型。

参考文献：

[1]David Hand 等.Principles of Data Mining(数据挖掘原理)[M].北京：机械工业出版社，2003.

[2]Datta A.Thomas H.The Cube Data Model:a Conceptual Model and Algebra for On-line Analytical Processing in Data Warehouses[J].Decision Support Systems,1999,27(3):289-301.

[3]刘蓉.利用数据仓库技术，完善综合电信管理决策系统[J].湖南省通信学会年会获奖论文，1999,11,(11):142-148.

[4]吕巍.移动通信运营商用户细分方法探讨[J].西部通信，2004,2,(2):32-37.

(责任编辑：董小玉)