第 36 卷 第 4 期 计 算 机 工 程 2010 年 2 月
Vol.36 No.4 Computer Engineering February 2010
·软件技术与数据库· 文章编号：1000—3428(2010)04—0071—03 文献标识码：A 中图分类号：TP311.13

基于对象立方体结构的类描述规则挖掘

赵秦怡，羊海潮
(大理学院数学与计算机学院，大理 671000)

摘 要：在类描述规则中，特征规则用于描述目标类中对象的特征，区分规则用于区分一个类及其对比类。研究基于对象立方体结构的类描述规则表示及其发现方法。通过实验验证该方法的可行性，得到用高层概念表示的类描述规则，该规则有助于用户对特定类进行识别。
关键词：对象立方体；类特征规则；类区分规则

Class Description Rule Mining Based on Object Cube Structure

ZHAO Qin-yi, YANG Hai-chao
(Institute of Mathematics and Computer, Dali University, Dali 671000)

【Abstract】 In class description rule, characteristic rule describes the characteristics of objects in target class, and discrimination rule distinguishes one class from other contrasting classes. It studies the class expression rule and its discovery method based on object cube construction. The feasibility of this method is validated by experiment and the class description rule expressed by high level conception is gained. The rule can help user to identify special classes.
【Key words】 object cube; class characteristic rule; class discrimination rule

近年来，随着面向对象技术的迅速发展以及面向对象数据库应用的日益广泛，基于面向对象数据库的数据挖掘成为新的研究方向之一[1-3]。通过将对象和类进行泛化，并构造对象立方体，可以发现类描述规则、分类规则、高层关联规则、例外规则、聚类规则等。类描述规则描述对象类的主要特征，是用户感兴趣的规则之一。类特征规则和类区分规则是类描述规则中较重要的 2 类规则。

1 对象立方体结构

对象立方体是类似于数据立方体[4-5]的 n 维数据模型，一个 n 维的数据立方体 $C[A_1, A_2, \cdots, A_n]$就是一个 n 维数据库，其中，$A_1, A_2, \cdots, A_n$ 是维数，每一维 A_i 表示关系的一个属性，共有$|A_i|+1$ 行，$|A_i|$是 A_i 维中不同值的个数，前$|A_i|$行是数据行，A_i 的每个不同值占一行，最后一行称为 *SUM* 行，用于存放上面所有行上 *COUNT* 列的总和，*COUNT* 是表决数[4]。

对象立方体是一个建立在泛化类上的多维数据库，以泛化类的一组已泛化属性作为对象立方体的维，而立方体的度量(各维空间)由一个或一组属性、聚集值或其他外延组成。其他外延可能是一个对象标识符的汇集或泛化对象其他特征的汇集。使用对象泛化技术可以把复杂对象的各种特征泛化成多个共享特征，构成多维数据库的共享维。面向对象数据库虽然含有大量具有复杂结构的对象、空间和多媒体数据以及类/子类层次结构和方法，但这些复杂结构可以泛化为简单和单一结构的数据，泛化后的结构和关系数据库结构很相似[6]。一个多维数据模型可以在泛化后的目标对象集上进行构造，因此，对象立方体可以通过基于维的属性泛化方法构造得到。

例 有如图 1 所示的类层次结构的 object 数据库，图 2 为其 address 属性概念层次树。泛化任务集如下：40 岁以上，有博士学位，在云南工作，开日产车。泛化任务如下：对该任务集中对象的姓名、住房面积、薪水、家庭地址进行泛化。对该任务用文献[7]中基于维的属性泛化算法进行泛化并构造相应的对象立方体，所得对象立方体如图 3 所示。

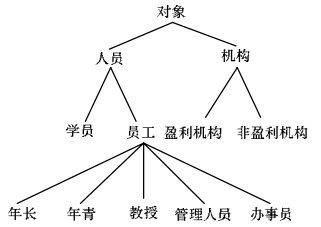

图 1 object 数据库类层次结构

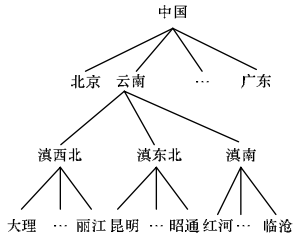

图 2 address 属性概念层次树

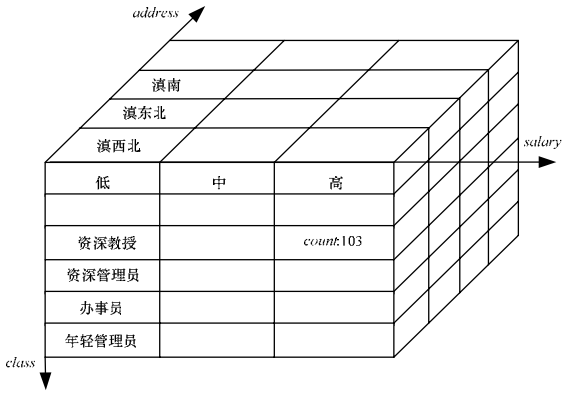

图 3 对象立方体

2 类特征规则及其挖掘

2.1 类特征规则

类的特征规则是指对满足任务相关对象集中所有或大多数对象特征的一个描述，一般表示成析取的逻辑规则形式[6]。有如下特征规则：

作者简介：赵秦怡(1973－)，女，副教授、硕士，主研方向：数据挖掘；羊海潮，硕士
收稿日期：2009-08-03 **E-mail：**dlzqu30@126.com

$(\forall)x\ target(x)\quad (house\_size(x)=\text{“小”}\ \vee salary(x)=\text{“低”})$

该规则说明了若 x 在目标对象集中，则 x 的住房面积是小的，且 x 的薪水是低的。特征规则中的概念应该是高层概念，代表的是目标对象集中最一般的行为。上述规则中的 $house\_size(x)=\text{“小”}$ 和 $salary(x)=\text{“低”}$ 即高层概念。

类的特征规则有 3 种表示方法：主类表，主类交叉表和带权重的量化规则。

2.2 类特征规则挖掘

将任务相关对象集进行泛化，进而构造得到相应的对象立方体，可以从中发现任务相关对象集的类特征规则。本文中类特征规则的挖掘均基于泛化的对象立方体结构。

(1)主类表

主类表是主泛化类的关系表格表示，是最简单直接的特征规则。表 1 是上例中任务相关对象集泛化后得到的对象立方体导出的主类关系表。主类表是任务相关对象集中最一般的特征表示，可以看成是一种特征规则表示。它在形式上和关系表非常相似，蕴含了泛化对象的逻辑规则及分布特征。如果主类中仅包含了一定数量的泛化对象，则通过主类表表现出来的特征规则往往很有效。

表 1 主泛化类表

class	*house_size*	*address*	*salary*	*count*
资深教授	大	滇东北	中等	64
资深教授	大	滇东北	高	103
…	…	…	…	…
资深管理员	小	滇南	低	49

(2)主泛化类交叉表

交叉表是主泛化类的另一种表示形式，可以通过主泛化类表绘制得到。绘制交叉表时，将每个属性作为交叉表的行或列，每个泛化的属性值作为表的一行或一列，表格的内容为满足行列条件的对象属性聚合值或满足条件对象的计数。表 2 是上例中泛化所得的主类表对应的主类交叉表。交叉表还可以绘制成若干子集交叉表，子集交叉表中仅含有任务相关属性集的一个子集，此类表更能反映子集中属性间的相关性，根据用户感兴趣的不同，属性子集可以绘制得到不同的子集交叉表。

表 2 主类交叉表

住房面积	地区	资深教授薪水				资深管理员薪水				总计
		低	中	高	总计	低	中	高	总计	
大	滇东北	0	64	103	167	0	16	20	36	203
	滇西北	2	27	46	75	1	8	9	18	93
	滇南	0	11	39	50	0	4	8	12	62
	总计	2	102	188	292	0	28	37	66	358
中	滇东北	1	42	20	63	0	12	1	13	76
	滇西北	0	21	62	83	0	3	18	21	104
	滇南	0	12	32	44	1	2	5	8	52
	总计	1	75	114	190	1	17	24	42	232
小	滇东北	4	10	0	14	8	0	0	8	22
	滇西北	11	4	2	17	2	9	0	11	28
	滇南	35	24	5	64	49	41	6	96	160
	总计	50	38	7	95	59	50	6	115	210
总计		53	215	309	577	61	95	67	223	800

(3)带权重的量化规则

主泛化类表或交叉表可以转化为带有权重的逻辑规则形式，其权重代表了规则中满足条件的对象在主类中出现的典型性，把这样的权重称为 $t_$权。令 q_a 为泛化后的条件，q_a 的 $t_$权是指覆盖了 q_a 的对象在初始对象集中所占的百分比，即

$$t\_\text{权} = count(q_a)/\sum_{i=1}^{n} count(q_i)$$

其中，n 是泛化结果对象集中的对象数；q_a 在 $q_1, q_1, \cdots, q_n$ 中。由上式可知，$t_$权的取值范围在 0~1 之间。带权重的泛化规则可以表示成如下 2 种形式：权重交叉表，量化的逻辑规则。

表 3 是上例中住房面积属性-类属性的权重交叉表，表 4 为住房面积-薪水的权重交叉表，权重交叉表可以通过对对象立方体或主泛化类表进行查询并计算得到，根据 $t_$权的定义可知其计算并不复杂。

表 3 住房面积-类属性的权重交叉表 (%)

住房面积	资深教授	资深管理员	总计
大	36.50	8.25	44.75
中	23.75	5.25	29.00
小	11.88	14.37	26.25
总计	72.12	27.88	100.00

表 4 住房面积-薪水的权重交叉表 (%)

住房面积	低	中	高	总计
大	0.38	16.25	28.12	44.75
中	0.25	11.50	17.25	29.00
小	13.62	11.00	1.63	26.25
总计	14.25	38.75	47.00	100.00

量化的特征规则描述了初始对象集中对象需满足的必要条件，规则中的条件并非是对象需满足的充分条件，因为满足该条件的对象可能出现在其他类中。量化规则可以表示成如下析取的逻辑规则形式：

$$(\forall)x\ \text{target\_class}(x)\rightarrow \text{condition}_1(x)[t{:}w_1]\vee \text{condition}_2(x)[t{:}w_2]\vee\cdots\vee \text{condition}_n(x)[t{:}w_n]$$

规则说明若对象 x 在初始对象类中，则 x 满足 condition_1 的概率为 w_1，依此类推，x 满足 condition_n 的概率为 w_n。用户可以提取大于某一 $t_$权阈值的量化规则，作为满足任务相关对象集中大多数对象的概括特征描述。量化的特征规则可以通过对对象立方体或主泛化类表进行条件查询、计算 $t_$权值而获得。

根据表 4 可得如下量化规则：

$\text{UBCJapCar}(x)\rightarrow(house\_size(x)=\text{“大”}\wedge salary(x)=\text{“中等”})$[16.25%]

$\vee(house\_size(x)=\text{“大”}\wedge salary(x)=\text{“高”})$[28.12%]

$\vee\cdots$

$\vee(house\_size(x)=\text{“小”}\wedge salary(x)=\text{“高”})$[1.63%]

上述规则说明，若对象 x 在目标集(驾驶日产车)中，则 x 住大面积房子且有中等收入的概率为 16.25%，x 住大面积房子且有高收入的概率为 28.12%，…，仅有 1.63%的概率为 x 住小面积房子且有高收入。

3 类区分规则及其挖掘

3.1 类区分规则

类的区分规则将一个任务相关对象集中的对象与其他对比类中的对象区别开[7]。例如，若将一个疾病区别与其他疾病，区分规则概括了该疾病区别于其他疾病的症状。目标类及对比类应该具有一定程度的相似性，即具有相同属性或维，否则区分规则的挖掘没有意义。在挖掘区分规则前需要先将目标类与对比类同步进行泛化，包括同时包含于目标类和对比类中的属性。能很好地将目标类区别于对比类的特征可能存在于对比类中，因此，在区分规则中需要有量化信息来给出区分规则的可信度，此类量化信息称为区分权重，记为 d-权，指一个特征发生在目标类中对比于发生在目标类及对比类中的概率。

令 q_a 是泛化后的条件，c_j 是目标类，q_a 的 d-权是指目标类中覆盖了 q_a 的对象数占目标类及对比类中覆盖了 q_a 的总对象数的百分比，即

d-权=$count(q_a \in c_j)/\sum_{i=1}^{k} q_a$

其中，$q_a \in C_i$；k 为目标类及对比类的总数。

由上式可知，d-权的取值范围在 0~1 之间，若 d-权值趋于 1，则说明该区分规则是一个较优的规则，若 d-权值趋于 0，则说明该区分规则不是一个优秀的区分规则。

利用 d-权值，区分规则提供了将目标类区别于对比类的量化标准。通常，量化的区分规则表示成如下逻辑规则形式：

$(\forall)x\ \ target\_class(x) \leftarrow condition_1(x)[d:w_1] \vee condition_2(x)[d:w_2] \vee \cdots \vee condition_n(x)[d:w_n]$

规则说明若 x 满足条件 i，则 x 在目标类中的概率为 w_i，i 值在 1~n 之间。

3.2 类区分规则挖掘

挖掘类区分规则的基本方法如下：根据区分规则得到目标对象集，即目标类及一个或若干个对比类，将目标类及对比类采用基于维的属性泛化方法进行同步泛化，得到主目标类及主对比类对象立方体。去除主目标类和对比类中相同的元组，d-权值将在剩余的元组中计算得到。计算元组的 d-权值，将 d-权值趋于 1 的元组析取在一起作为区分规则进行输出。

在上例的任务相关对象集中，有如下目标类：在高校工作，住房面积=“大”。其对比类如下：在高校工作，住房面积=“小”。发现该目标类区别于对比类的区分规则。在图 3 所示的对象立方体中进行查询，汇集出目标类和对比类，计算出相应的 d-权值，可得如下用户最感兴趣的区分规则：

If salary=“low” then house_size=“small”[97.32%]

If salary=“high” and address=“滇东北” then house_size=“big” [100%]

If salary=“low” and address=“滇南” then house_size=“small” [100%]

4 结束语

在面向对象数据库中，对任务相关对象集(对象类)进行描述是用户感兴趣的知识之一。类特征规则对满足任务相关对象集中全部或多数对象的特征进行描述，类区分规则将目标类中的对象和对比类中的对象区分开。在表示规则时，高层概念的规则较底层概念的规则更具代表性和典型性。基于泛化对象立方体的描述规则是高层概念表示的规则。

参考文献

[1] Ham J, Nishio S, Kawano H. Knowledge Discovery in Object-oriented and Active Databases[M]. [S. l.]: IOS Press, 1994.

[2] Kim W. Introduction to Object-oriented Database[M]. [S. l.]: MIT Press, 1990.

[3] Ester M, Kriegel H P, Xu Xiaowei. Knowledge Discovery in Large Spatial Databases: Focusing Techniques for Efficient Class Identification[C]//Proc. of the 4th Int'l Symp. on Large Spatial Databases. Los Alamitos, USA: IEEE CS Press, 1995: 67-82.

[4] 樊 博, 李海刚, 孟庆国. 空间数据立方体的建模方法研究[J]. 计算机工程, 2007, 33(8): 1-2.

[5] 刘亚波, 刘大有, 高 滢, 等. 基于数据立方体的属性核计算方法[J]. 计算机工程, 2008, 34(20): 46-48.

[6] Han Jiawei. 数据挖掘概念与技术[M]. 范 明, 孟小峰, 译. 北京: 机械工业出版社, 2004.

[7] Han Jiawei, Nishio S, Kawano H, et al. Generalization-based Data Mining in Object-oriented Databases Using an Object-cube Model[J]. Data and Knowledge Engineering, 1998, 25(1/2): 55-97.

编辑 陈 晖

(上接第 70 页)

际项目的风险管理工作中。在实践中发现，将完整项目分解为加工单元，以降低复杂度的方法具有可行性，在相同网络节点数量规模下，其结构比贝叶斯网络方法易于建立和维护。利用该方法可以在复杂项目中快速计算和观察风险影响的估算数据。

在元素成功概率的递推计算方面，本文方法与贝叶斯网络风险分析方法[5]的明显不同在于：(1)本文方法采用概率加法和乘法公式的自由组合计算公式来组织元素间的概率递推关系，而不使用贝叶斯公式，计算中不需要输入条件概率分布的真值表，因此，很大程度上降低了网络节点数据的维护难度。例如，针对 50 个元素节点，若按每个元素有 4 个前驱父节点来计算，则本文方法需要输入 50 组概率即可，而贝叶斯网络方法至少需要 50×2^4=800 组。本文方法在建模时不需要考虑全局所有元素节点的条件独立性问题，在贝叶斯网络方法中则必须考虑该问题。(2)本文方法可以得到基于元素和基于过程(即加工单元)的 2 套拓扑结构，基于加工单元的拓扑图对于复杂项目来说，更容易理解且便于沟通和实用，而贝叶斯网络方法仅能得到一种基于节点的拓扑结构。

表 3 为 3 个项目的部分实践数据统计，其资金规模都在 50 万~80 万人民币之间，前 2 个是商业网站建设项目，另一个是金融行业专用软件项目。项目 A 作为对照，未使用此工具，项目 B 和项目 C 使用此工具后，风险监管数量有所增加，但风险信息维护和沟通效率得到明显加强。实践结果表明，RiskManager 工具能把风险影响较直观地显示出来，具有计算快速和调整便利灵活的优势，有利于项目干系人之间的风险管理信息沟通。该工具有助于风险影响估算、风险分析的经验积累和辅助风险管理者的决策。

表 3 RiskManager 工具实践数据

风险管理情况	项目 A*	项目 B	项目 C
风险监管数量	10	20	20
风险统计报表数/(份·周$^{-1}$)	2	6	6
平均风险源更新工作量/(人天)	4.0	2.5	2.0
平均风险沟通工作量/(人天)	5.0	3.0	3.0

4 结束语

本文提出一种基于软件过程的项目风险管理辅助工具。其中，加工单元设计的细粒度判断以及元素成功概率的设取必须由经验丰富的使用者来执行，以获得良好的风险管理效果。下一步工作将从操作便利性和数据挖掘深入两方面着手，继续完善工具的开发，并计划将元素成功概率细化到项目范围、进度、质量等其他维度。

参考文献

[1] 项目管理协会. 项目管理知识体系指南[M]. 3 版. 北京: 电子工业出版社, 2004.

[2] Wallace L, Keil M. Software Project Risks and Their Effect on Outcomes[J]. Communication of the ACM, 2004, 47(4): 68-69.

[3] Boehm B, Huang Liguo. Value-based Software Engineering: Reinventing “Earned Value” Monitoring and Control[J]. Software Engineering Notes, 2003, 28(2): 1-7.

[4] 杨振明. 概率论[M]. 北京: 科学出版社, 1999.

[5] 冯 楠, 李敏强, 寇纪淞, 等. 基于贝叶斯网络的软件项目风险管理模型[J]. 计算机工程, 2007, 33(7): 41-43.

编辑 陈 晖