计算机科学2004Vol. 31№. 4

空间数据概念树提升算法研究*)

郑志强　吴春明　余建桥
(西南农业大学信息学院　重庆400716)

摘　要　本文提出了空间数据概念树提升算法，该算法对基于规则的面向属性的概念树提升算法作了改进，主要是引入可性度因子对规则进行修正，并实行主概念属性一次性泛化。整个算法无需回溯。

关键词　空间数据，泛化，可信度

A Study of Clustering Algorithm on Spatial Data Concept Tree

ZHENG Zhi-Qiang　WU Chun-Ming　YU Jian-Qiao
(Information College, Southwest Agricultural University, Chongqing 400716)

Abstract　This paper has put forward the clustering algorithm on spatial data mining concept tree. The algorithm, based upon Rule Based Attribute-oriented concept tree, reaches the rule conclusion by introducing the creditability factor modification rule, generalizing the concept only once and using statistic result to correct error. The algorithm does not need back tracking.

Keywords　Spatial data, Generalization, Creditability

1　引言

空间数据是指描述"空间实体"的空间位置特征和专题属性的数据。近年来，海量数据增长越来越快，数据的内容及表现形式都发生了重大变化。传统的数据挖掘(Data Mining)和知识发现(Knowledge Discovery in Database，简称KDD)算法多针对于关系数据库，并不完全适应空间数据挖掘。其中，JiaWei Han等人提出的基于规则的面向属性的概念树提升算法[1,2](Rule Based Attribute-Oriented，简称RBAO)比较典型地实现了非空间数据库数据知识挖掘；通过属性泛化和回溯两遍扫描提升概念，具有较好效果。但该算法需要回溯，效率不是特别高，泛化结果与统计值之间并无必然联系。本文提出，对空间数据来说，一方面泛化过程除了定性处理之外还需要定量处理；另一方面规则意味着知识，对规则的再加工可以提高泛化效率。因此引入可信度因子建立泛化评价系统，不仅只需一次扫描即可完成泛化，而且对泛化结果能够验证和修正，具有较高效率。

2　基于规则的面向属性的概念树算法(RBAO)

面向属性的概念树提升算法就是以领域专家提出的概念分层次逐步泛化最后得出一阶谓词逻辑表示的规则的算法。

依赖规则的面向属性的概念树提升算法是利用规则来进行属性概念的泛化。

RBAO算法分为两步进行：

(1)按照属性阈值和泛化规则，提升概念树，形成主关系表(Prime-relation)。即应用统计原则和专家知识，确定每一属性的泛化阈值，换句话说合理区分不同值之间的分段。并根据泛化规则形成主关系。

(2)利用回溯算法，形成最终关系(Final-relation)表

主关系表中某些属性应进一步泛化，直到关系的规模小于泛化关系阈值。泛化过程中，逐步去掉了某些属性。这些属性有可能在规则上又对其他更高层概念属性有着影响，为了体现出这些属性在挖掘中的价值，RBAO算法应用回溯算法恢复丢失的信息。即①把主关系中的元组回溯到它们的源集。在初始关系中加入一个虚拟属性(覆盖元组标识符)来记录对应的覆盖元组。②选择某些属性提升到更高级别，然后比较合并广义元组。③合并元组映射回主关系并分裂主关系中的元组。加强主关系中具有相同虚拟属性值的所有元组映射为主关系中对应的覆盖元组。主关系中的元组因此分裂为若干个元组，统计结果同时被调整。④继续泛化，得出最终规则结论表。

3　改进的空间数据概念树提升算法

RBAO算法依赖规则构建了关于属性的概念树，并通过对概念的提升实现了知识发现。但为了准确体现出属性间的关系，必须进行回溯且效率不高。另外，结论的可信度也难以验证。下面我们通过改进规则，重构概念树进行泛化，并引入可信度因子对泛化结果进行评价和修正。整个算法无需回溯，大大提高了效率和泛化结果的可靠性。

首先引入规则表示方法：$A(x) \wedge B(x) \rightarrow C(x)$，即对于元组x，如果不同于属性A的属性B满足条件，则属性A所对应的概念就可以泛化到高一层次的概念层C。逐步泛化直到最高层次概念即可得出结论。

对上述泛化规则进行改进，重写为

$$(A(x)=c) \leftarrow (A(x)=a)\&B_1(x)\&B_2(x)\&\cdots\&B_n(x)$$

即如果元组x符合$A(x)=a$，同时满足属性$B_1, B_2, \cdots, B_n$，则属性A从概念a上升到概念c。又如果存在另一规则$(A(x)=d) \leftarrow (A(x)=c)\&D_1(x)\&D_2(x)\cdots\&D_n(x)$

*)重庆市科委项目(20038035)资助。**郑志强**　硕士，讲师，主要研究方向是人工智能、软件工程等。**吴春明**　硕士，讲师，主要研究方向是人工智能、数据库系统等。**余建桥**　博士，教授，主要研究方向是人工智能、数据库系统、软件工程等。

则可以推导出

$(A(x)=d)\leftarrow(A(x)=c)\&D_1(x)\&D_2(x)\cdots\&D_n(x)\leftarrow(A(x)=a)\&B_1(x)\&B_2(x)\cdots\&B_n(x)\&D_1(x)\&D_2(x)\cdots\&D_n(x)$[3] (1)

设属性 A 的概念层从低到高依次为 $a_1,a_2,\cdots,a_n$,(设 a_n 为最高概念层),基于属性的规则集合为 R,则对于任意规则 r,有

$(A(x)=a_i)\leftarrow(A(x)=a_k)\ \hat{}\ r\ \hat{}\ \cdots \quad (k<i)$

由式(1)可得 $(A(x)=a_n)\leftarrow(A(x)=a_{n-1})\ \hat{}\ r\leftarrow\cdots\leftarrow(A(x)=a_1)\ \hat{}\ r_1\ \hat{}\ r_2\cdots$

因此规则集合中的所有规则都可以进行修改,形成如下表示形式:

$(A(x)=a_n)\leftarrow(A(x)=a_1)\ \hat{}\ r_1\ \hat{}\ r_2\cdots$ (2)

所以对于任一规则 r,通过与其他规则共同作用,总可以在泛化过程中发挥作用。由式(2)可以看出:在概念树提升过程中,可以抛弃属性 A 概念提升的中间过程,直接从最低层泛化到最高层。而且在规则集合中已经涵盖了中间的规则过程,不会造成中间信息丢失。

在前述规则中所有规则都描述了各自的事实,没有轻重之分。在概念提升过程中起的作用也是一致的。最后推理出的结果与统计值可能有较大出入。这种将规则一概化的做法是与客观事实不相符合的。例如,商场附近有地铁、车站的概率为90%,没有的概率为10%。将其引入规则表示中,简单地表示为{商场}←{交通便利}就不能真实反映事实。事实上,这里数字是至关重要的。

现在引入可信度因子修改规则。设总量为1,上述规则的可信度为0.9(虽然是概率统计数字,但将其设为可信度是合理的)。具体表示为:{商场}←{交通便利}(0.9)。这样在一定程度上准确地表述了事实。将可信度因子引入泛化过程,通过它与统计值进行比较,最终验证推理结果。

设 a_{11},a_{12} 为 $A(x)=a_1$ 的阈值,a_{21},a_{22} 为 $A(x)=a_2$ 的阈值,b_1,b_2 为属性 B 的阈值

设 $A(x)=a_{11}$ 的可信度为 α_{11},$A(x)=a_{12}$ 的可信度为 α_{12}

$B(x)=b_1$ 的可信度为 β_1,$B(x)=b_2$ 的可信度为 β_2

则对于规则 $r_1(A(x)=a_{21})\leftarrow(A(x)=a_{11})\ \hat{}\ (B(x)=b_1)$ 的可信度为 $\alpha_{11}\times\beta_1$

同理对于规则 $r_2(A(x)=a_{22})\leftarrow(A(x)=a_{12})\ \hat{}\ (B(x)=b_2)$ 的可信度为 $\alpha_{12}\times\beta_2$

这样,在规则修改中,同时引入可信度因子来标记该规则的可信度。在概念提升过程中,虽然属性有可能丢弃掉了,但可信度因子通过上述处理方法或加权平均法得以继续发挥影响。这样,在最终结论中,通过可信度因子和统计结果可以衡量结论是否准确、合理,是否应该抛弃。

下面是改进的空间数据概念树提升算法的具体步骤:

(1)引入可信度因子,修改规则;

(2)直接泛化主概念属性到最高层次,同时变化可信度;

(3)泛化其他相关属性和独立属性(独立属性是指与任一规则都不相关联的属性);

(4)比较可信度与统计结果,修正推理结果,并去掉不合理的结论。

4 算法验证

一个关于城市商场地理特征的数据库,包括商场周边一平方公里范围内的人口密度、月平均收入、环境(有无公园湖泊等)、交通、净收益等信息,目标是净收益的特征规则。如净收益与环境有无必然联系等。

图1是由领域专家提供的概念树。

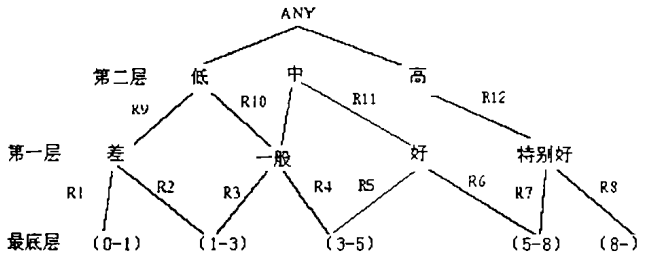

图1 净收益的规则依赖概念图

根据对净收益这一属性的概念层的分划,需要从最底层逐步泛化到最高层。下面是一些规则:

R1:{0-1}→差;

R2:{1-3}^{人口密度>30000}→差;

……

R12:{好}→中;

R13:{特别好}→高;

现在,引入可信度因子修改所有的规则,形成如下的规则集合:

{净收益=高}←{净收入为8万元以上}(0.1)

{净收益=高}←{净收入为5~8万元}&{人口密度<30000}(0.05)

{净收益=中}←{净收入为5~8万元}&{人口密度>30000}(0.2)

{净收益=中}←{净收入为3~5万元}&{人口密度<30000}(0.06)

{净收益=中}←{净收入为1~3万元}&{人均月收入<2000}&{人口密度<30000}(0.01)

{净收益=中}←{净收入为3~5万元}&{人口密度>30000}&{人均月收入<2000}(0.048)

{净收益=低}←{净收入为0~1万元}(0.1)

{净收益=低}←{净收入为1~3万元}&{人口密度>30000}(0.2)

{净收益=低}←{净收入为1~3万元}&{人均月收入>2000}&{人口密度<30000}(0.04)

{净收益=低}←{净收入为3~5万元}&{人口密度>30000}&{人均月收入>2000}(0.192)

对于净收入、人口密度、人均月收入的概率统计值见下表1。

表1

	净收入					人均月收入		人口密度	
概率值	0-1	1-3	3-5	5-8	8-	高	低	稠密	稀疏
	0.1	0.25	0.3	0.25	0.1	0.8	0.2	0.8	0.2

综观规则集合,主概念属性净收益是受到净收入、月平均收入、人口密度等因素的制约和影响的。概念提升的实质就是要将净收益概念上升到高、中、低三种概念层次,对其他属性进行概念提升和适当的属性丢弃。首先将净收益属性直接泛化到最高层次,然后再泛化其他的相关属性(这里指月平均收入、人口密度),最后泛化独立属性。

表2是提升净收益属性到最高层次的泛化结果。

表2

净收益	人口密度	人均月收入	可信度	统计值
高	?	?	0.1	50
高	<30000	?	0.05	35
中	>30000	?	0.2	28
中	<30000	<2000	0.01	33
…	…	…	…	…
低	>30000	>2000	0.192	20

对人口密度、人均月收入等属性设定其到泛化阈值，再按照上表泛化各属性，并分别修正可信度和统计值，得出规则结论。

从结果上看，本算法在运算效率上，优于 RBAO 算法大约30%～40%，特别是利用可信度因子丢弃了一些在 RBAO 算法中根本无法判别其是否病态的结论，收到了良好的效果。

结束语 本算法引入可信度因子并实行主概念属性一次性泛化来实施概念树提升，不仅解决了算法回溯的问题，而且对推理结果进行了有效验证。因而其效率和准确性都有了较大提升。周生炳等人提出的无回溯的 RBAO 算法，更多地考虑了属性间的依赖关系，但空间数据库数据之间并无明显的属性依赖关系。由规则得出的属性依赖往往是不准确的。本算法主要引入可信度因子并在泛化过程中施以变化，对于解决空间数据推理问题具有较强的适应性和效率。作者利用本算法在地理信息系统中进行概念提升获得了较好的效果。

参 考 文 献

1 Stefanovic N, Han J, Koperski K. Object-Based Selective Materialization for Efficient Implementation of Spatial Data Cubes. IEEE Transactions on Knowledge and Data Engineering, 2000,12(6)

2 Han J,Cai Y D,Cercone N. Knowledge discovery in databases: An attribute-oriented Approach. In: Proc. of 18th Intel Conf. on Very Large Data Bases, Vancouver, Aug. 1992. 547～559

3 周生炳，张钹，成栋. 基于规则面向属性的数据库归纳的无回溯算法. 软件学报，1999,10(7)

4 Xu X, Ester M, Kriegel H-P, Sander J. A Distribution-Based Clustering Algorithm for Mining in Large Spatial Databases. In: Proc. 14th Int. Conf. on Data Engineering (ICDE'98), Orlando, FL, 1998

5 Han J, Stefanovic N, Koperski K. Selective Materialization: An Efficient Method for Spatial Data Cube Construction. In: Proc. 1998 Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD'98), Melbourne, Australia, April 1998

6 Lu W,Han J. Information Associated Join Index for Spatial Range Search. International Journal of Geographical Information Systems, 1995,9(3):221～249

7 Ester M, Kriegel H-P, Sander J. Spatial Data Mining: A Database Approach. In: Proc. 5th Symp. on Spatial Databases, Berlin, Germany, 1997

8 Han J, Huang Y, Cercone N,et al. Intelligent query answering by knowledge discovery techniques. IEEE Transactions on Knowledge and Data Engineering, 1996,8(3):373～390

9 Carter C L, Hamilton H J. Performance improvement in the implementation of DBLEARN: [Technical Report 94-5]. Department of Computer Science, University of Regina, Sask., Canada, Jan. 1994

（上接第116页）

将基于 Apriori 算法实现的 OLAP 中的关联规则挖掘同本文中的算法从性能上比较来看：首先原型算法的性能不同。FP-增长算法只要遍历数据库2次，Apriori 算法则要根据最终生成的模式长度决定，最终生成的频繁模式越长，遍历数据库的次数越多。Apriori 算法需要频繁的连接、候选、产生频繁项的步骤，并且最小支持度越低，候选集越大，开销也越庞大，FP-增长算法则在第二次遍历数据库后，将整个数据库压缩到 FP 树上，通过 FP 树挖掘一次性地产生所有的频繁模式，省去了候选过程。其次本文提出的恰当的数据结构提高了性能。在单维中对于稀疏数据我们提出了表2的存储结构，减少了计算频繁1-项集时的读取单元数和比较次数。在混合维关联规则挖掘方面，同事辛燕采用 Apriori 算法实现了 OLAP 中的混合维关联规则挖掘，其中要判断连接项进行维内连接还是维间连接，根据维内连接还是维间连接再采用不同的判定条件决定是否可以进行连接，本文中我们在表3的基础上采用 FP-增长算法则巧妙地避免了这一系列的复杂比较，并将其与多维关联挖掘进行了合并。由此可以看到，采用本文中介绍的算法和数据结构可以大大提高性能。

除了文中介绍的几种关联规则挖掘，利用 OLAP 还可容易实现多层次关联规则挖掘：依据 OLAP 数据结构，首先获得感兴趣的任何层次的数据，然后调用规则挖掘算法。对于基于约束的关联规则挖掘，本文只是基于事实值对生成模式进行了一定的约束，实际情况中，还可以限定模式的前件、后件以及模式长度等，在进一步的研究中，我们将在此方面进行更多的研究。

参 考 文 献

1 Chaudhuri S, Dayal U. An overview of data warehousing and OLAP technology. ACM SIGMOD Record, 1997,26:65～74

2 Han Jiawei. Towards On-Line Analytical Mining in Large Databases. 1998

3 Han Jiawei, Chee S H S, Chiang J Y. Issues for On-Line Analytical Mining of Data Warehouses. 1998

4 Hu Xiaohua, Cercone N. An OLAM Framwork for Web Usage Mining and Business Intelligence Reporting. 2002 IEEE

5 Fabris C C, Freitas A A. Incorporating Deviation-Detection Functionality into the OLAP Paradigm. 2001

6 Zhu Hua. Online analytical mining of association rules. 1998

7 Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. In: Proc. of the ACM SIGMOD Conf. on Management of data, 1993. 207～216

8 Houtsma M. A run Swam i. Set-oriented mining of association rules [A]: [Research Report RJ 9567]. IBM Almaden Research Center [C]. San Jose, California: [s. n.] 1993

9 Han Jiawei, Pei Jian, Yin Yiwen. Mining Frequent Patterns without Candidate Generation. In: Proc. of the ACM Int. Conf. on Management of Data, Dallas, TX, May 2000