利用空间数据采掘技术挖掘气候模式

华东船舶工业学院电子与信息系 江苏 镇江 刘同明 刘 伟

探讨利用空间数据采掘技术挖掘气候模式的方法。空间数据挖掘系统的体系结构和挖掘方法必须考虑空间对象同时具有空间数据和与其有关的非空间数据这一特殊性，提出两种空间数据挖掘策略，并根据这些策略开发了镇江气候模式挖掘系统。本文讨论的挖掘模型可用于其他空间数据挖掘系统。

本文探讨空间数据的挖掘方法，并依据这些方法初步开发了一个镇江地区气候模式发现系统(ZJCPMiner)，测试表明，该系统能够发现用户感兴趣的气象模式。

空间数据挖掘的模型

图1表示一个空间数据挖掘模型。用户通过控制器控制知识发现的每一步。空间数据和非空间数据的概念层次结构以及数据库信息等背景知识存放在知识库中。数据库接口优化用户查询，并汇集与学习任务有关的数据。聚焦模块判定哪一部分数据对模式识别有用。模式提取模块用于发现规则和模式，同时把统计方法、机器学习或者数据采掘技术与计算几何算法组合起来，共同完成知识发现任务。评价模块评价所提取的模式，去除冗余知识。通过控制器交互，提供反馈信息，所有的发现都送给用户验证。

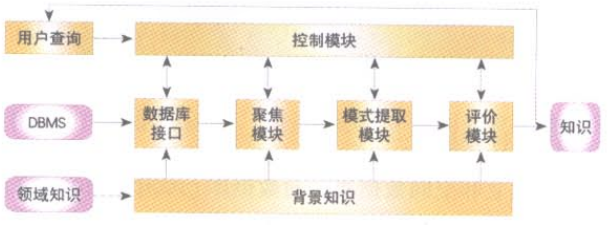

图 1 空间数据采掘模型

空间数据采掘的方法

1. 空间数据挖掘方法的基础

国内外学者主要针对关系数据挖掘提出的许多方法和算法，为空间数据挖掘奠定了基础。例如，数据泛化(generalization)技术以及基于泛化的面向属性归纳(AOI: Attribute-Oriented Induction)学习方法等。但在SDM中引入关系数据挖掘方法时，要充分考虑空间对象的特点。一个空间对象同时具有空间和非空间数据，那么不仅要考虑对它们同时泛化，也还有一个主次问题。AOI能够发现空间数据和非空间数据之间的联系，但对于涉及不同主题地图信息的系统，要求AOI方法能够分析不同主题地图上的不同空间特征之间的关系。数据聚类也是如此，要考虑空间区域及空间距离度量准则。总之，可以借鉴或引用关系数据挖掘的概念和方法，但必须加以改进和扩充。

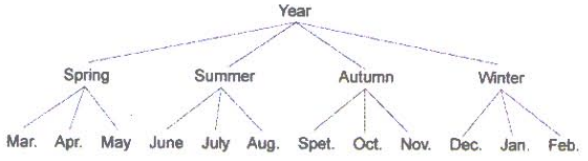

图 2 “季节”概念层次结构

2. 四点假设

假设一:知识发现过程有可用的背景知识，由领域专家提供或从数据库自动生成。背景知识以概念层次结构(树结构、表结构或集合形式)描述，图2是一个树结构的例子，表1是表结构的例子。

© 中国科学院软件研究所 http://www.c-s-a.org.cn

表 1 降雨量概念层次 单位: 英寸

非常干燥	干燥	比较干燥	一般	比较湿润	湿润	非常湿润
[0, 0.1]	(0.1, 0.3]	(0.3, 1]	(1.0, 1.2]	(1.2, 2.0]	(2.0, 5.0]	5.0及以上

假设二:数据挖掘的任务是提取一般的特征规则或互联关系,学习过程由一个用户请求触发。用户的学习请求用类SQL语言描述。

例1:给出镇江地区各气象测报站采集的气象数据集(每月的气温和降雨量),学习任务是发现镇江地区1996年夏季的气候模式。该学习请求用类SQL描述为:

EXTRACT characteristic rule

FROM precipitation-map, temperaturre-map

WHERE area="镇江" AND period="夏季" AND year=1998

IN RELEVANCE TO region AND precipitation and temperature

其中,“夏季”是根据图2所示的“季节”概念层次由月份(6、7、8月)泛化得到的概念。

假设三: 空间数据库由空间对象的空间数据和与其有关的非空间数据组成。非空间数据以二维表形式存放在关系数据库中。空间数据按主题存储,每一张主题图包含了空间对象的空间特征,主题图用光栅方法表示(亦可用矢量方法)。图3表示镇江地区气象观测站分布的主题图,它同时给出了对象的空间层次。

图 3 镇江地区气象观测站分布情况主题图

假设四:空间数据库的容量足够大,数据足够多,且稳定可靠。

3. 两种空间数据泛化策略

由于空间对象同时带有空间和非空间数据,所以,泛化时需要两种概念层次:主题概念层次和空间概念层次,在泛化一种数据时,同时相应地调整另一种数据。不同的应用将决定首先对哪部分数据泛化,即采用空间数据控制的泛化还是非空间数据控制的泛化。

(1)非空间数据控制的泛化。这种泛化策略以非空间数据为主,同时调整空间数据。

例2:给出季节概念层次(图1)、降雨量概念层次(表1)和气象测报站分布图(图3),以及1998年的降雨量数据(表2)。

表 2 镇江地区1996年降雨量数据 单位: 英寸

Station	Jan.	Feb.	Mar	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
大港	0.37	1.36	1.99	2.20	1.47	5.55	4.91	3.01	2.73	4.19	1.06	3.11
丹阳	0.6	1.45	2.01	2.4	1.9	3.7	3.8	3.1	2.4	4.5	1.3	2.4
句容	0.86	1.23	1.87	2.32	1, 6	4.9	4.0	2.8	2.5	3.7	1.5	2.3
扬中	1.2	1.74	2.15	2.68	2.51	6.4	5.8	3.7	3.1	4.3	1.5	2.6
…	…	…	…	…	…	…	…	…	…	…	…	…

假定学习任务是发现镇江地区1998年夏季各县市的降雨量模式,用类SQL语言表示如下:

EXTRACT region

FROM precipitation-map

WHERE city="镇江" AND peiod="夏季" AND year=1998

IN RELEVANCE TO precipitation and region

泛化过程如下:

①建立非空间数据集。对数据库precipitation执行SQL查询,提取符合要求的降雨量记录,形成目标数据集(非空间数据集)。本例中period=“夏季”是一条泛化数据,由季节概念层次得到。

②在非空间数据集中执行面向属性的泛化。首先根据图1的季节概念层次把6、7、8三个月泛化成“夏季”;

© 中国科学院软件研究所 http://www.c-s-a.org.cn

计算这三个月的平均降雨量，作为降雨量的泛化值(见表3)，但应根据表1降雨量概念层次作进一步泛化，最后结果如表3所示。

表 3 泛化后的降雨量数据

Station	June	July	Aug.	Avg.	泛化值
大港	5.55	4.91	3.01	4.49	非常湿润
丹阳	3.7	3.8	3.1	3.53	湿润
句容	4.9	4.0	2.8	3.9	湿润
扬中	6.4	5.8	3.7	5.3	湿润
…	…	…	…	…	…

表 4 降雨量泛化信息

区域	降雨量
沿江(Ⅰ)	非常湿润
丹徒(Ⅱ)	湿润
茅山(Ⅲ)	比较湿润
丹阳(Ⅳ)	湿润

注意，在泛化过程中，要在泛化的非空间数据入口中保存指向空间对象(测报站)的指针。

③执行空间泛化。当把非空间数据泛化到与用户学习请求相应的概念层次后，合并具有相同泛化属性值的邻近区域。在此过程中，用邻近函数或空间谓词(如closeto(x，y)等)发现邻近区域，并把空间对象分布到合并后的几个区域中。还可能用到近似方法，即，在该空间区域中如果只有极少数地区支持某个泛化属性值，则可以忽略这个空间区域。

学习结果如表4(ZJCPMiner系统同时在电子地图上显示降雨量分布特征，这里省略了)。

(2)空间数据控制的泛化。也可以首先根据空间层次信息泛化空间数据进行泛化，其过程与非空间数据控制的泛化相似，其中，空间概念层次可根据以下方法得到:①利用空间数据的语义，如行政区域;②空间对象聚类;③利用空间索引结构，如R树、四叉树、八叉树等。

表 5 气温概念层次

非常寒冷	寒冷	比较寒冷	温暖	比较炎热	炎 热	非常炎热
-10及以下	(-10,0)	(0,10)	(10,20)	(20，30)	(30，35)	35及以上

表 6 丹阳市1998年夏季气温数据

Station	June	July	Aug	Avg
司徒	21	29	33	27.6
延陵	28	34	34	32
珥陵	28	31	35	31.3
皇塘	28	30	33	30.3
吕城	24	28	34	28.6
云阳	26	25	35	28.6
访仙	28	29	31	29.3
新桥	27	32	31	30

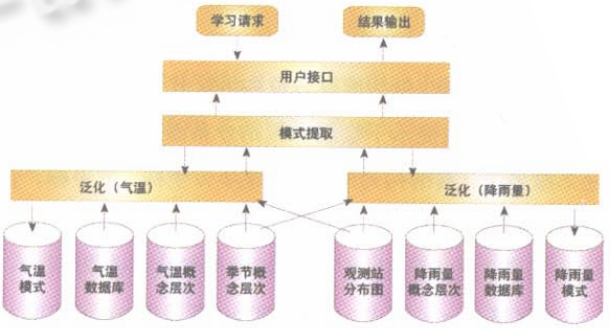

图 4 系统结构框图

ZJCPMiner系统的开发

1. 系统组成

基于关系数据库Virsual Foxpro建立“镇江地区气候数据库”，存放近年来的气温和降雨量数据。每一年度的气候数据(气温或降雨量)存储在一个表中。气象观测站分布图中存储着各观测站的空间信息(地理位置和区域代码)。概念层次树由专业气象人员提供。学习任务是发现镇江地区某一时期的气候模式。系统结构框图如图4。应用程序用Visual Basic编制。

2. 具体实现

(1)概念树的数据结构的设计。概念层次结构是系统频繁访问的数据信息，其数据结构影响系统的工作效率。概念层次结构以树型结构为好。系统一般从叶结点出发按“层”存取概念树信息，另外，本系统中概念树的叶结点都是用上界和下界定义的一个数据范围，所以叶结点的数据结构如下:

```
Type Concept
    DownLimit As Single
```


© 中国科学院软件研究所 http://www.c-s-a.org.cn

```
    UpLimit As Single
    Value_1 As String * 10   '叶结点的父结点值
    Value_2 As String * 5    '叶结点祖先结点值
End Type
```

整个概念层次的数据信息可存储在叶结点数组中。例如:

```
Dim TemperatureConceptTree(N)
For i=1 To N
  输入TemperatureConceptTree(i).DownLimt
  输入TemperatureConceptTree(i).UpLimt
  输入TemperatureConceptTree(i).Value_1
  输入TemperatureConceptTree(i).Value_2
Next I
```

(2)气象观测站分布图的设计。分布图中存储着各观测站的空间信息,包括地理方位和地区代码。系统初始化时,用户在系统提示下绘制观测站分布图,并将信息存储在二维表中,其结构为:观测站名,X坐标,Y坐标,管辖范围面积。

(3)模式的表达形式。系统提取的模式表示为:

<区域>→<模式><支持度><覆盖度><可信度>

例如,

西部地区→比较寒冷 95% 90% 89%

其中,支持度、覆盖度和可信度三个参数用于描述模式的可靠性。由于种种原因,测报数据存在一定的误差,从这些数据中提取的模式的准确性和可信性必然受到影响,故使用"可信度"参数表示模式的可靠性,其取值范围为[0, 1)。

从目标数据集合中提取的每一个模式都对应集合中的一个记录子集,该子集合中记录的个数与整个数据集中记录数的比值,就是这个模式的覆盖度,取值范围[0, 100%],所有模式的覆盖度之和应当是"100%"。如果子集中的记录与模式匹配,称它完全支持该模式。完全支持相应模式的记录数与记录子集中记录个数的比值,就是这个模式的支持度。用户可以根据以上参数对提取的模式进行评价。■

参考文献

1 Krzysztof K, Jiawei H and Junas A. Mining Knowledge in Geographical Data. http://db.cs.cfu.ca/GeoMiner/surves

2 Ng R and Han J. Efficient and effetive clustering method for spatial data mining. In Proceedings of 1994 International Conference Very Large Data Bases(Step. 12-15, Santiago, Chile). Morgan Kaufmann, San Francisco, CA, 1994, p.144~155.

3 Matheus C. J, Chan P.K. Systems for Knowledge Discovery in Databases. IEEE Trans. Knowledge and Data Engineering 5, 5(Oct. 1993), p.903~913

© 中国科学院软件研究所 http://www.c-s-a.org.cn