ISSN1000-9825 Journal of Software 软件学报 2000,11(5):673~678

基于 Rough Set 的空间数据分类方法*

石云 孙玉芳 左春

(中国科学院软件研究所 北京 100080)

E-mail: shiyun263@263.net

摘要 近来,数据采掘的研究已从关系型和事务型数据库扩展到空间数据库.空间数据采掘是一个很有发展前景的领域,其中空间数据分类的研究尚处在起步阶段.该文分析和比较了现有的几个空间数据分类方法的利和弊,提出利用 Rough Set 的三阶段空间分类过程.实验结果表明,该算法对于解决包含不完整空间信息的问题是有效的.

关键词 Rough Set,分类,数据采掘,空间数据,约简.

中图法分类号 TP181

近来,数据采掘的研究已从关系型和事务型数据库扩展到空间数据库,空间数据采掘是一个很有发展前景的领域.目前,已陆续有一些对空间关联算法[1]、空间聚类算法[2]、泛化的空间描述[3]、空间聚类的特征描述[4]等方面的研究,但这些研究还缺乏对空间对象分类算法的有效模型.虽然统计学和机器学习领域针对关系型数据已有一系列分类方法[5~7],但是地理数据包含空间对象和这些对象的非空间描述,空间分类的标准不仅包含对象的非空间属性,还包含分类对象与其他对象间的空间关系,因此具有与普通分类不同的难点.

空间数据分类的研究目前尚处在起步阶段.Fayyad 等人使用决策树方法对大约 3TB 的卫星图像中的星系对象进行分类[8].数据图像先经低层图像处理系统 FOCAS[8]的预处理,选择欲分类对象并生成图像要素、面积、密度、方向等基本属性,通过天文学家对训练数据集中的对象进行分类,利用决策树算法可找到规则集合.该方法处理的是图像数据库,专用于天文学应用,不适用于采用矢量数据格式的空间数据的分析.

Ester 等人提出一种基于 ID3(interactive dichotomizer 3)算法[6]的空间对象分类方法[9],采用邻域图,分类标准基于分类对象的非空间属性以及描述分类对象与其邻近位置相关对象间空间关系的属性、谓词和函数.该方法的缺点是不分析邻近对象非空间属性的聚合值,而在实际中,如果一个对象在其邻近区域内某属性的聚合值与另一个对象邻近若干个区域内对应属性的聚合值相同,则这两个对象的属性就应视为类似.同时,对该算法也没有进行相关性分析,因此可能会导致生成低质量的决策树.

Koperski 和 Han 对 Ester 等人算法[9]中相应的问题进行了改进[10],降低了计算时间复杂度.但是,基于决策树的分类算法不适合处理带有不完整信息的问题.空间数据分类标准中包含数据间的空间关系,从某个训练数据集来讲,空间属性极有可能缺失.如果输入数据出现了不一致、噪声等情况,决策树算法可能会造成误分,从而严重影响决策树算法的预测准确度.采用决策树空间分类算法不能很好地体现地理空间关系对于分类的影响.

本文提出基于 Rough Set[11]的空间数据分类方法,较好地解决了上述问题.其三阶段方法为:第 1 步,提取空间谓词;第 2 步,根据属性重要度从空间谓词集合中选取空间决策属性,或从属性集合中选取非空间决策属性;第 3 步,对条件属性集合进行约简,发现 Rough Set 分类规则.实验结果表明,在训练数据集中缺失某些属性的情况下,基于 Rough Set 算法生成的规则在灵敏度方面比采用决策树算法具有更好的性能.

---

* 本文研究得到国家 863 高科技项目基金(No.863-306-ZD-07-4)资助.作者石云,女,1972 年生,博士,主要研究领域为数据库知识发现,空间数据采掘.孙玉芳,1947 年生,研究员,博士生导师,主要研究领域为大型数据库和网络应用系统,操作系统,中文信息处理.左春,1959 年生,研究员,主要研究领域为大型数据库和网络应用系统,数据库知识发现.

本文通讯联系人:石云,北京 100080,中国科学院软件研究所

本文 1999-02-04 收到原稿,1999-05-24 收到修改稿

1 采用 Rough Set 进行空间分类

空间分类的目标是为了发现分类规则.参与分类处理的标签可以有以下 4 种类型:(1) 对象的非空间属性;(2) 具有非空间值的空间相关属性;(3) 空间谓词;(4) 空间函数.空间函数为空间谓词的组合,因此,在下文中如无特别说明,均将空间谓词和空间函数统称为空间谓词.本文的分类标准沿用文献[9,10]中的规定,即基于(1) 分类对象的非空间属性;(2) 描述分类对象和位于分类对象邻近位置的对象间空间关系的属性和谓词.

例 1:图 1 为一简单的示意性地图,根据对象 O_i 邻近对象的描述和 O_i 的非空间属性,对地图上的 5 个对象 $O_i(i=1,2,3,4,5)$进行分类.利用缓冲区 A,B,C,D,E 来表述 O_i 邻近对象的属性.表 1 列出围绕对象 O_i 所建立的缓冲区中相关属性的非空间描述.

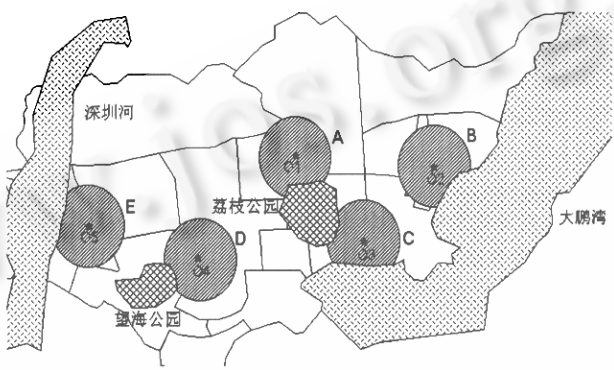

Fig. 1　Example map
图 1　示例地图

Table 1　Non-Spatial description of relevant attributes in the buffer
表 1　缓冲区中相关属性的非空间描述

Buffer①	Population②	Average income③	Ratio of crimes④(%)
A	50,000	2,550	0.2
B	35,000	5,000	0.3
C	55,000	2,550	1
D	66,000	3,250	0.5
E	80,000	4,550	0.4

①缓冲区,②人口,③平均收入,④犯罪率.

缓冲区代表对分类对象的类标签属性有影响的区域.它既包含用于从专题地图中查找聚合信息的缓冲区,也包含用于决定空间谓词的缓冲区.缓冲区的大小和形状选择是一个很重要的问题,可以使用不同的准则来决定缓冲区的形状.缓冲区可能是环状区域、Voronoi 图式等[12].因为数据中没有其他可用信息以辅助决定缓冲区的大小和形状,因此在实验中我们采用如图 1 所示的等距离缓冲区.

1.1 空间谓词的提取

如表 2 所示,空间谓词可以用于描述分类对象.为提高算法效率,描述首先应该泛化.根据预先给定的概念层次建立泛化规则,泛化结果见表 3.

查找空间谓词的过程时间复杂度很高.为加快该过程,可采用两阶段处理方法:首先进行粗略计算,然后只对那些有希望的模式进行细化计算.第 1 步,先查找采样对象的粗略描述,可使用最小边界矩形法(MBRs)来查找 coarse_g_close_to 谓词(即两个对象的最小边界矩形位于具体的距离范围之内).第 2 步,采用机器学习方法来抽取相关谓词或函数[13,14].在实验中,我们采用 Relief 算法[14],其算法描述如算法 1 所示.Relief 算法采用最近相邻方法来查找相关的谓词:对采样中的每个对象 s,查找两个与它最邻近的对象,其中一个与对象 s 属于同类(nearest hit),另外一个与对象 s 属于不同的类(nearest miss).基于最近相邻的对象的描述,对谓词的权重作修改.如果邻近对象和对象 s 同属一类,并具有相同的谓词值,则该谓词的权重增加;如果邻近对象和对象 s 同属

一类,但不具有相同的谓词值,则该谓词的权重减少;如果邻近对象和对象 s 不属于同一类,却具有相同的谓词值,则该谓词的权重减少;如果邻近对象和对象 s 不属于同一类,也不具有相同的谓词值,则该谓词的权重增加.可以看出,相关谓词的权重为正值,而不相关谓词的权重期望值为 0.最终,仅有那些权重大于预定义的阈值的谓词用于分类.相应阈值可基于统计理论来设定[14].

Table 2 Descriptions of classified objects

表 2 分类对象的描述

O_i	Spatial predicate①
O_1	Close_to (x,荔枝公园)
O_2	Close_to (x,大鹏湾)
O_3	Close_to (x,荔枝公园)
	Close_to (x,大鹏湾)
O_4	Close_to (x,望海公园)
O_5	Close_to (x,深圳河)

①空间谓词.

Table 3 Generalized descriptions of classified objects

表 3 分类对象的泛化描述

O_i	Spatial predicate①
O_1	Close_to (x,公园)
O_2	Close_to (x,水域)
O_3	Close_to (x,公园)
	Close_to (x,水域)
O_4	Close_to (x,公园)
O_5	Close_to (x,水域)

①空间谓词.

算法 1. Relief 算法

```
Procedure Relief (Predicate_weight, k)
  FOR j:=1 TO max_predicate DO
    Predicate_weight[j]:=0;
  FOR sample_i:=1 TO min(200,k) DO /* size of the sample is set to min(200,k) */
    nearest_hit:=FIND_NEAREST_HIT(sample_i); /* find nearest sample from the same class */
    nearest_miss:=FIND_NEAREST_MISS(sample_i); /* find nearest sample from the different
                                                  class */
    FOR j:=1 TO max_predicate DO
      Predicate_weight[j]:=Predicate_weight[j]-diff(sample_i,nearest_hit,j)
                                              +diff(sample_i,nearest_miss,j);
  FOR j:=1 TO max_predicate DO
    IF Predicate_weight[j]>min(200,k)*threshold/√min(200,k)
      THEN Predicate_relevant[j]=TRUE
      ELSE Predicate_relevant[j]=FALSE;
Function diff(sample_i,nearest_hit_or_miss,j)
/* for symbolic attributes */
    IF sample_i[j]=nearest_hit_or_miss[j] THEN return (0)
      ELSE return (1).
```

采用相关性分析计算相关谓词,可以减小信息表中属性集的大小,提高 Rough Set 算法的效率和分类的准确率.

1.2 选取决策属性

在空间数据分类中,决策属性有空间决策属性和非空间决策属性两种.选取适合的决策属性对分类规则的建立十分重要.非空间决策属性的选取有较为成熟的方法,此处不再赘述.对于空间决策属性,因为空间谓词是在线分析提取出来的,很难判断出哪个可以作为决策属性,本文利用属性重要度来选择最佳空间决策属性.

空间谓词查找完毕后,生成一个谓词集合 PA,通过比较各个谓词的属性重要度,选取属性重要度最大的空间谓词作为空间决策属性.算法描述如下.

算法2. 选取空间决策性的算法

DO

max_measure=0;

FOR 属性 $a\in PA$ DO /* 选取一个属性重要度值最大的属性 */

measure=AttributeSignificance(a);

IF (measure>max_measure)

max_measure=measure;

best_attr=a.

1.3 对条件属性集合进行约简并发现 Rough Set 分类规则

选取好决策属性后,将剩下的空间谓词集合与属性集合合并,作为条件属性集合,组成信息表 S. 信息表 S 中的属性数目一般较多,可先进行预处理,去掉多余属性,以便提高发现效率,降低错误率,并排除噪声干扰.

对 S 采用基于差别矩阵的属性约简策略[15],如算法3所示. 由于生成一个决策表的所有约简或计算出属性数目最少的约简是NP难问题,因而该策略对最小约简是不完备的,但是根据文献[15]的结果,该策略对多数情况所获得的属性约简集合是最小的.

算法3. 属性约简算法

令 M 是信息表 S 的差别矩阵,M 的元素表示为 A_{ij},它是 S 中第 i 个实例与第 j 个实例有差别的所有属性的集合. $A=\{a_1,a_2,\ldots,a_n\}$ 是 S 所有属性的集合. $A_{ij}\subseteq A$,$a_k\in A_{ij}$. C_0 是 S 的所有约简的交集,称为 M 的核(C_0 可能为空).

令 $p(a_k)$ 是 M 中属性 a_k 的属性频率函数,它定义为属性 a_k 在 M 中出现的次数. 属性约简策略如下:

令 $R=C_0$,

(1) $Q=\{A_{ij}:A_{ij}\cap R\neq\varnothing,i\neq j,i,j=1,2,\ldots,n\}$,$M=M-Q$,$B=A-R$;

(2) 对所有 $a_k\in B$,计算在 M 中的 $p(a_k)$ 并且令 $p(a_q)=\max_k\{p(a_k)\}$;

(3) $R\Leftarrow R\cup\{a_q\}$;

(4) 重复上述过程,直到 $M=\varnothing$.

则集合 R 是信息表 S 的一个属性约简.

属性约简后,就可以抽取空间分类规则,其算法描述如算法4所示. 其中,计算例子 E 的约简的算法即是算法3中表述的内容.

算法4. 分类规则学习算法

输入:一组训练例子集 ES.

输出:规则集 RS.

过程:

(1) 初始化:$RS=\varnothing$;

(2) FOR 训练集 ES 中的每个例子 E DO

reduct=ComputeReduct(E,ES)

/* 计算例子 E 的约简 */

rule=GenerateRule(reduct,E)

/* 抽取出 E 中与约简对应的属性值形成规则 */

$RS=RS+\{$rule$\}$

如上所述,采用 Rough Set 方法进行空间分类有两个特点:(1) 采用属性约简方法去掉多余属性和空间谓词;(2) 利用属性重要度选择最佳空间决策属性,提高采掘的针对性.

2 分类算法

2.1 算法描述

输入:

(1) 空间数据库包含:

(a) 对象集合

(b) 具有非空间属性的其他空间对象

(2) 数据采掘查询指定:

(a) 在分类处理中使用的对象

(b) 相关属性和谓词

(3) 非空间概念层次集合

输出:Rough Set 分类规则

处理过程:

第1步.提取空间谓词,包含以下过程:

(1) 采集查询中指定的数据集合 S,包含分类对象和用于描述的对象的集合.

(2) 对于 S 中的样本空间对象 O_i:

(a) 使用粗略的谓词、函数和属性建立描述所有对象的谓词集合.

(b) 基于概念层次执行谓词集合的泛化.

(c) 采用 Relief 算法查找与分类任务有关的粗略的谓词、函数和属性.

(3) 采用相关性分析,建立描述分类对象的空间谓词的集合.

(4) 基于概念层次执行谓词集合的泛化.

第2步.选取决策属性.

第3步.对条件属性集合进行约简,并发现 Rough Set 分类规则.

2.2 算法的性能分析

本文对由 MapInfo Professional 4.1地理信息系统生成的实验数据集(包含有3 000多个对象)进行了测试.基于谓词 g_close_to,分别测试采用文献[10]中 ID3决策树算法与采用本文 Rough Set 算法的分类质量和算法的执行时间,用以评估算法的质量和效率.谓词 g_close_to 代表分类对象和位于距离阈值内的对象间的空间关系.从每个类标签集合中选取400个对象用于训练集,余下的50个对象用于测试.实验结果见表4.

Table 4 Experimental results

表4 实验结果

Number of predicates①	Time[s](ID3)②	Time[s](Rough set)③	Quality(ID3)④(%)	Quality(Rough set)⑤(%)
20	105	150	60	80
12	76	98	75	86
3	12	17	90	100

①谓词数量,②时间[秒](决策树方法),③时间[秒](Rough Set 方法),④质量(决策树方法),⑤质量(Rough Set 方法).

实验结果表明,采用 Rough Set 方法进行空间分类与文献[10]所采用的 ID3决策树算法相比,时间复杂度要高一些.但 ID3算法在建造决策树时对噪声较为敏感,如果输入数据不完整或出现噪声,就会严重影响决策树算法的预测准确度.空间数据分类在分类标签中使用的空间谓词往往带有不确定性和不完整信息,数据噪声出现的几率较大,ID3算法难以克服这些问题.利用 Rough Set 方法进行预处理,可以去掉多余的属性和谓词,提高发现效率,降低错误率.在训练数据集中缺失某些属性的情况下,基于 Rough Set 的算法生成的规则在灵敏度方面具有更好的性能.另外,采用属性重要度概念决定空间决策属性,增强了分析的针对性.因此,采用 Rough Set 方法进行空间分类可提高分类处理的质量.

3 结束语

采用 Rough Set 方法进行空间对象分类能够较好地反映空间和非空间数据之间的关系,为利用邻近区域中基于非空间属性的聚合值来对空间对象进行分类提供了可行性.我们在中国人民保险公司深圳分公司保险业务分析软件中已部分实现了该算法.今后将对空间对象间的关系进行更深层次的研究,探索如何将关联分析与分类规则相结合,以便更好地进行空间数据采掘.

参考文献

1 Koperski K, Han J W. Discovery of spatial association rules in geographic information databases. In: Egenhofer M J ed. Proceedings of the 4th International Symposium on Spatial Databases (SSD'95). Advances in Spatial Databases. Berlin: Springer-Verlag, 1995. 47～66

2 Ester M, Kriegel H P, Sander J *et al*. A density-based algorithm for discovering clusters in large spatial databases. In: Simoudis E, Han J W, Fayyad U M eds. Proceedings of the 2nd International Conference on Data Mining (KDD-96). Portland: Oregon, 1996. 226～231

3 Lu W, Han J W, Ooi B C *et al*. Discovery of general knowlecge in large spatial databases. In: Proceedings of Far East Workshop on Geographic Information Systems. Singapore: World Scientific, 1993. 275～289

4 Ng R T, Yu Y. Discovering strong, common and discriminating characteristics of clusters from thematic maps. In: Proceedings of the 11th Annual Symposium on Geographic Information Systems. 1997. 392～394

5 Fayyad U M, Piatetsky-Shapiro G, Smyth P *et al*. Advances in Knowledge Discovery and Data Mining. Menlo Park, CA: AAAI/MIT Press, 1996

6 Quinlan J R. Induction of decision trees. Machine Learning, 1986,(1):81～106

7 Safavian S R, Landgrebe D. A survey of decision tree classifier technology. IEEE Transactions on Systems, Man and Cybernetics, 1991,21(3):660～674

8 Fayyad U M, Djorgovski S G, Weir N *et al*. Automating the analysis and cataloging of sky surveys. In: Advances in Knowledge Discovery and Data Mining. Menlo Park, CA: AAAI/MIT Press, 1996

9 Ester M, Kriegel H P, Sander J. Spatial data mining: a database approach. In: Scholl M, Voisard A eds. Proceedings of the International Symposium on Large Spatial Databases (SSD'97). Berlin, New York: Springer Verlag, 1997. 47～66

10 Koperski K, Han J W, Stefanovic N. An efficient two step method for classification of spatial data. In: Poiker T ed. Proceedings of the 1998 International Symposium on Spatial Data Handling (SDH'98). Vancouver, BC, 1998

11 Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning About Data. Amsterdam, North-Holland: Kluwer Academic Publishers, 1992

12 Peterson K. A trade area primer. Business Geographics, 1997,5(9):18～21

13 Wetteschereck D, Aha D W, Mohri T. A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. Artificial Intelligence Review, 1997,(10):1～37

14 Kira K, Rendell L A. The feature selection problem: traditional methods and a new algorithm. In: Proceedings of the 10th National Conference on Artificial Intelligence (AAAI-92). Cambridge, MA: MIT Press, 1992. 129～134

15 Wang Jue, Wang Ren, Miao Duo-qian *et al*. Data enriching based on rough set theory. Chinese Journal of Computers, 1998,21(5):393～400
(王珏,王任,苗夺谦等. 基于 Rough Set 理论的"数据浓缩". 计算机学报,1998,21(5):393～400)

Spatial Data Classification Based on Rough Set

SHI Yun　SUN Yu-fang　ZUO Chun

(*Institute of Software　The Chinese Academy of Sciences　Beijing*　100080)

Abstract　Recent studies have extended the scope of data mining from relational and transactional databases to spatial databases. Spatial data mining is a promising field, where the research work on spatial data classification is still in its initial stage. In this paper, the advantages and the disadvantages of several existing methods of spatial data classification are compared first. Then an effective three-step method, which is based on the rough set theory for spatial data classification, is proposed. The validity of this algorithm to the problem of incomplete spatial information is verified by pertinent experimental results.

Key words　Rough set, classification, data mining, spatial data, reduction.