基于 co-location 模式的空间分类算法

赵秦怡[1] 王丽珍[2] 罗桂兰[1]
[1](大理大学数学与计算机学院 云南 大理 671003)
[2](云南大学信息学院 云南 昆明 650091)

摘 要 在特定的空间分类任务中,对象的类别和自身属性相关较小,和近邻对象的空间特征相关较大,用传统的空间分类方法并不适用。提出一种基于 co-location 模式的空间分类挖掘算法。算法挖掘含不同类别特征的空间 co-location 模式,转化为分类规则,获得兴趣度较高的分类规则集。分类阶段先查询待分类对象的空间近邻,概化为空间特征,挑选适应的分类规则进行分类。实验结果表明这是一种高效的空间分类算法。

关键词 空间分类 空间 co-location 模式 空间数据挖掘

中图分类号 TP311 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2018.03.043

A SPATIAL CLASSIFICATION ALGORITHM BASED ON CO-LOCATION PATTERNS

Zhao Qinyi[1] Wang Lizhen[2] Luo Guilan[1]
[1](*School of Math and Computer Science, Dali University, Dali 671003, Yunnan, China*)
[2](*School of Information and Engineering, Yunnan University, Kunming 650091, Yunnan, China*)

Abstract In a particular spatial classification task, the category of an object is less related to its own attributes and has a greater correlation with the spatial features of a neighbouring object, which is not applicable using the traditional method of spatial classification. This paper presented a co-location-based spatial classification mining algorithm. The algorithm mined spatial co-location patterns with different categories of features into the classification rules to obtain the classification rules set with high interest. In the classification stage, the spatial neighbours of the objects to be classified were inquired, and the classification was characterized by generalization of spatial features and classification. Experimental results showed that this was an efficient spatial classification algorithm.

Keywords Spatial classification Spatial co-location patterns Spatial data mining

0 引 言

空间分类是指对空间对象分类时,除了要考虑待分类对象的非空间属性对分类结果的影响[1],还要考虑其空间邻接对象对分类结果的影响[2]。Fayyad 等[2]提出了一种空间决策树分类方法,使用决策树对卫星图像中的星系对象进行分类。Ester 等[3]提出一种基于 ID3 算法的空间分类方法,分类标准基于待分类对象的非空间属性及空间属性、谓词和函数。Koperski 等[4]对 Ester 等的算法进行了改进,降低了算法的时间复杂度。Shekhar 等[5]提出了一种基于粗糙集的空间分类方法,采用空间谓词对空间关系进行泛化,再使用粗糙集对数据进行分类。

空间 co-location 模式挖掘是指发现一组空间特征集合 c,c 中空间特征的实例在地理空间中频繁出现[6]。基于全连接的 Join-based 算法[7]将 Apriori 算法思想引入了空间 co-location 模式挖掘中,利用特征实例间的邻近关系挖掘 co-location 模式。部分连接的 partial-join 算法[7]把连续空间中的实例分割为不相交的团,并且通过邻近关系的断点保持这些团之间的关系。文献[8]提出了一种基于星型邻居扩展的无连接 joinless 算法,算法不需要通过连接实例来产生 co-location 模式表实例。针对产生的表实例开销大的问题,

收稿日期:2017-06-16。国家自然科学基金项目(61661001)。**赵秦怡**,副教授,主研领域:空间数据挖掘算法。**王丽珍**,教授。**罗桂兰**,副教授。

王丽珍等[9-11]提出了基于前缀树的 co-location 模式挖掘方法。

在满足某种 co-location 模式的空间中,一些具有特定特征空间对象的出现意味着另一种特定特征空间对象的出现[12]。如半湿润常绿阔叶林生长的地方80%有兰类植物的生长,有尼罗河鳄鱼的地方85%会有埃及鸻。特定环境下的 co-location 模式中含有空间分类所需的分类规则,本文提出了基于 co-location 模式的空间分类算法。

1 基本概念

1) 空间 co-location 模式[12]:一个空间 co-location 模式是一组空间特征的集合 c,这些空间特征的实例在地理空间中频繁地出现,其中 $c \subseteq F$。

例:{半湿润常绿阔叶林,兰类植物}是一个 co-location模式。

2) 参与度($PR(c)$)[12]:参与度是指衡量 co-location模式 c 的频繁性所使用的支持度标准,它的取值是 co-location 模式 c 的所有空间特征参与率(PR值)中的最小值,记为 $PI(c)$。

参与率[12]记为 $PR(c,f_i)$,是特征 f_i 的实例在 co-location 模式 c 的所有实例中不重复出现的个数与 f_i 总实例个数的比率,其计算式如下:

$$PR(c,f_i) = \frac{|\Pi_{f_i}(table\_instance(c))|}{|table\_instance(\{f_i\})|} \quad (1)$$

3) 类别特征

定义 1 在空间分类问题中,待分类对象的类标号属性值域记为 C,由 C 中的每一个元素值定义的空间特征称为类别特征。

例 1:在植物生长区域的分类任务中,类属性"是否有兰类植物生长"值域为 $\{Yes,No\}$,可将该分类任务的类别特征集定义为 $\{CS_1,CS_2\}$,CS_1 代表特征"有兰类植物生长",CS_2 代表特征"没有兰类植物生长"。

4) 与分类任务相关的空间 co-location 模式

定义 2 与分类任务相关的空间 co-location 模式是指一个含有类别特征的空间 co-location 模式(记为 CB co-location)。一个 CB co-location 模式的表实例满足最小参与度阈值。

例 2:在某分类任务中,类别特征集 $CS = \{CS_1, CS_2, CS_3\}$,与该分类问题相关的空间属性集 $F = \{A, B, C, D\}$,模式 $\{CS_1, A\}$、$\{CS_2, B\}$、$\{CS_1, A, B\}$、$\{CS_2, C, D\}$、$\{CS_3, B, C\}$ 含某一类别特征,参与度大于给定的最小参与度阈值,它们是 CB co-location 模式。而模式 $\{A,B\}$、$\{B,C,D\}$ 不含类别特征,不论参与度为多少,它们不是 CB co-location 模式。

5) 规则 $A \rightarrow B$ 的置信度:指出现特征 A 的情况下,特征 A、B 共同出现的概率,是衡量规则可信度的有效性度量,记为 $C_{onfidence}(A \rightarrow B)$,计算式表示为:

$$C_{onfidence}(A \rightarrow B) = \frac{P(A \cup B)}{P(A)} \quad (2)$$

2 基于 co-location 模式的空间分类算法

2.1 算法思想

在特定的空间分类任务中,空间对象的类别和自身非空间属性相关较小,和空间近邻对象的特征相关较大[12-13],用传统的空间分类方法来进行分类并不适用。如加油站选址分类任务主要考察近邻对象是否具备如高速路、交通关口、交通流量等特征;移动服务运营商针对不同地区设置相应的移动服务需求模式;广告商在某类人群聚集区域放置不同种类的广告。在前述的分类问题中,目标对象的类别由近邻对象的特征决定,和自身非空间属性的相关关系可以忽略。若将分类任务进行泛化,即将近邻对象及目标对象的类别进行泛化,得近邻特征及类别特征,前述的分类任务可描述为:在特定空间模式下,近邻特征集的不同决定了空间目标对象类别的不同,即特定空间中的 co-location 模式含空间分类所需的分类规则。

基于 co-location 模式的空间分类是指利用含类别特征的空间 co-location 模式对空间对象进行分类。算法需在特征实例集中挖掘所有含任一类别特征的 CB co-location 模式。第一步,在训练阶段根据类标号属性的值域得到若干相应的空间类别特征 $CS_1 \sim CS_n$,n 为分类问题中类属性的取值个数。确定与分类问题相关的空间特征集 A,该项工作可由领域专家辅助完成。由类别特征 CS_1 和 A 构成空间特征集 AR,在 AR 的实例集中挖掘含有类别特征 CS_1 的 CB co-location 模式;由类别特征 CS_2 和 A 构成空间特征集 AR,挖掘含类别特征 CS_2 的 CB co-location 模式;重复该过程,直至挖掘出含类别特征 CS_n 的 CB co-location 模式。将上述所有的 CB co-location 模式归并,构成目标模式集 CB。第二步,生成 CB 对应的分类规则集 R,规则的后件为类别特征,计算规则的置信度。由于 CB co-location 模式均为频繁模式,故规则的支持度(兴趣度标准)在算法中不再度量。第三步,分类阶段在空间近邻集中查询出待分类对象的近邻对象集,将其概化为近邻空间特征

集[14],运算出该特征集的各子集,找出分类规则集 R 中所有包含任一子集的分类规则,由其中支持度最大的分类规则中规则后件(类别特征)决定待分类对象的类别。

例3:某空间分类问题中,训练集类别特征集 $CS=\{CS_1,CS_2\}$,与分类任务相关空间特征集 $A=\{A,B,C,D,E\}$,参与度阈值 70%。各特征实例为:$CS_1.1 \sim CS_1.3, CS_2.1 \sim CS_2.3, A.1 \sim A.4, B.1 \sim B.4, C.1 \sim C.4, D.1 \sim D.4, E.1 \sim E.4$。特征实例间的近邻关系如图1所示。待分类空间对象 O_1、O_2、O_3,含分类对象的空间近邻关系集略。

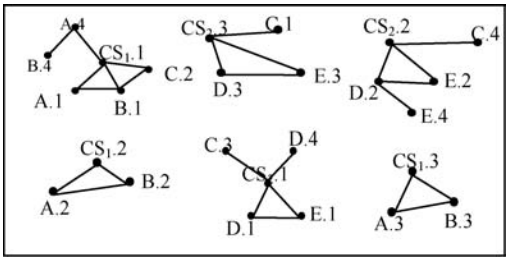

图1　例3中特征实例间的近邻关系图

基于 co-location 模式的分类过程如下:

1) 生成空间特征集 $AR=\{CS_1,A,B,C,D,E\}$,挖掘含类别特征 CS_1 的 CB co-location 模式。所得含 CS_1 的频繁模式如表1所示。

表1　含 CS_1 的 CB co-location 模式

模式	CS_1	A	CS_1	B	CS_1	A	B
表实例	$CS_1.1$	A.1	$CS_1.1$	B.2	$CS_1.1$	A.1	B.2
	$CS_1.2$	A.2	$CS_1.2$	B.3	$CS_1.2$	A.2	B.3
	$CS_1.3$	A.3	$CS_1.3$	B.1	$CS_1.3$	A.3	B.3
	$CS_1.1$	A.4					
参与率	3/3	4/4	3/3	3/4	3/3	3/4	3/4
参与度	1		3/4		3/4		

得到目标模式集合 $CB=\{\{CS_1,A\},\{CS_1,B\},\{CS_1,A,B\}\}$。

2) 生成空间特征集 $AR=\{CS_2,A,B,C,D,E\}$,挖掘含类别特征 CS_2 的 CB co-location 模式。所得含特征 CS_2 的频繁模式如表2所示。

表2　含 CS_2 的 CB co-location 模式

模式	CS_2	C	CS_2	D	CS_2	E	CS_2	D	E
表实例	$CS_2.1$	C.3	$CS_2.1$	D.4	$CS_2.1$	E.1	$CS_2.1$	D.1	E.1
	$CS_2.2$	C.4	$CS_2.2$	D.2	$CS_2.2$	E.2	$CS_2.2$	D.2	E.2
	$CS_2.3$	C.1	$CS_2.3$	D.3	$CS_2.3$	E.3	$CS_2.3$	D.3	E.3
参与率	3/3	3/4	3/3	3/4	3/3	3/4	3/3	3/4	3/4
参与度	3/4		3/4		3/4		3/4		

得到目标模式集合 $CB=\{\{CS_1,A\},\{CS_1,B\},\{CS_1,A,B\},\{CS_2,C\},\{CS_2,D\},\{CS_2,E\},\{CS_2,D,E\}\}$。

3) 由集合 CB 导出分类规则并计算规则的置信度。得规则集 $R=\{A\rightarrow CS_1(100\%),B\rightarrow CS_1(75\%),(A,B)\rightarrow CS_1(75\%),C\rightarrow CS_2(75\%),D\rightarrow CS_2(75\%),E\rightarrow CS_2(75\%),(D,E)\rightarrow CS_2(75\%)\}$。

4) 在空间数据库中查询待分类空间对象 O_1、O_2、O_3 的近邻对象,将近邻对象概化为空间特征,得到 O_1 的近邻空间特征集 $\{A,B,E\}$,O_2 的近邻特征集 $\{D,E\}$,O_3 的近邻特征集 $\{A,B,C,D,E\}$。对象 O_1 的近邻特征集的所有子集为 $\{A\}$,$\{B\}$,$\{E\}$,$\{A,B\}$,$\{A,E\}$,$\{B,E\}$,$\{A,B,E\}$,集合 R 中的规则 $A\rightarrow CS_1(100\%)$,$B\rightarrow CS_1(75\%)$,$(A,B)\rightarrow CS_1(75\%)$ 包含了子集 $\{A\}$,$\{B\}$,$\{A,B\}$,其他规则不包含 O_1 的任意子集。故对象 O_1 的类别由规则 $A\rightarrow CS_1(100\%)$,$B\rightarrow CS_1(75\%)$,$(A,B)\rightarrow CS_1(75\%)$ 决定。根据算法,由参与度最大的规则 $A\rightarrow CS_1(100\%)$ 得到 O_1 的类别为 CS_1。O_2 和 O_3 的分类过程同上,得 O_2 的类别为 CS_2,O_3 的类别为 CS_1。

空间 co-location 模式在挖掘时,将 k 阶特征集(前 n 项相同)相连接而得到 $k+1$ 阶特征集。由此可知在挖掘与类别相关的空间 co-location 模式时,若 k 阶 co-location 模式中含类别特征,则 $k+1$ 阶 co-location 模式也含类别特征。在挖掘2阶 co-location 模式时,只需考虑含类别特征的二阶模式集,将类别特征在空间特征集中排在第一位,由模式组合方法,即前 $n-1$ 个特征相同进行组合,可知挖掘出的更高阶 co-location 模式也一定含类别特征。这样,在剪枝阶段避免了大量剪枝,算法的计算复杂度得到了有效降低。

2.2　算法描述

算法1　基于 co-location 模式的空间分类算法

输入:实例近邻关系集 T,类别特征集 S,分类任务相关空间特征集 A,空间类别特征集 CS。

输出:R,C_label。

算法中所用符号说明见表3。

表3　符号说明

名称	含义
T	特征实例近邻关系集
S	类别特征实例集
A	分类任务相关空间特征集
CS	类别特征集
$minf$	最小参与度阈值
AR	含任一类别特征的分类任务相关空间特征集
CB	与分类任务相关的 CB co-location 模式集
R	CB 集中模式对应的分类规则集

续表 3

名称	含义
O	待分类对象
N	待分类对象 O 的近邻对象集
NA	由 N 概化得到的对象 O 近邻特征集
OB	与 O 相关的分类规则集
CY	OB 中参与度最大的分类规则
C_label	O 的类标号属性值(输出)

算法描述:

Step1 数据预处理:在实例近邻集 T 中删除不含任一类别特征实例的完全独立连接。

Step2 由集合 CS 和 A 生成含类别特征 CS_i 的分类任务相关空间特征集$(CS_i, A_1, A_2, \cdots, A_n) \rightarrow AR$。

Step3 挖掘满足最小参与度阈值 *minf* 并且含特征 CS_i 的 CB co-location$\rightarrow CB$。

Step4 i ++,若 $i \leqslant n$,转 Step2。(这样所有含 $CS_1 \sim CS_n$ 中某一特征的 CB co-location 均挖掘出)。

Step5 生成形如 $X \rightarrow CS_i$ 的分类规则集 R:对 CB 中的每一个 co-location 模式,模式中的第一个特征(CS_i)作为规则的后件,模式中第 2 个特征起始的所有特征作为规则的前件,扫描近邻实例集,计算规则的置信度。

Step6 查询对象 O 的近邻对象集 N,将 N 概化为近邻特征集 NA,计算 NA 的所有子集。

Step7 对每一子集 NA_i,若 NA_i 等于 R 中某一规则的前件,则将该规则 $\rightarrow OB$。

Step8 挑选出 OB 中置信度最大的规则 CY,CY 的后件 $\rightarrow C\_label$。

算法在挖掘二阶模式时,只需挖掘含类别特征的二阶模式,故不需要将特征集中所有特征两两组合。算法中候选模式生成阶段的 SQL 伪代码如下:

```
if k = =1
 //在初始属性集中将类别特征 CS_i 放在元素 f_1 的位置上
{ for q = f_2 to f_n
        select CS_i, q - > insert into C_2
}
else
{ forall co-location p ∈ C_k
        forall co-location q ∈ C_k
            insert into C_{k+1}
            select p.f_1, p.f_2, ⋯, p.f_k, q.f_{k+1}
            where p.f_1 = q.f_1, ⋯, p.f_k = q.f_k, p.f_{k+1} < q.f_{k+1}
}
```

3 算法分析及实验

算法的时间复杂度集中在二阶表实例的搜索及分类规则支持度计算阶段,若特征实例数越多,需要的运行时间越长。其次,若分类问题的类标号属性值越多,需要挖掘的含类别特征的分类规则增加,算法的运行时间随之增加。设有 n 个类别,m 个特征,每个特征 k 个实例,算法时间耗费主要在星型模型生成、二阶频繁模式生成、模式连接及表实例搜索、规则参与度计算等阶段,算法总的时间复杂度为 $O(n(k^2 + 2^{m+1}k))$。

通过实验对算法的分类准确率和运行时间进行了评价。实验环境:intel core(TM)i7-7500U 的 CPU,2.7 GHz主频,8 GB 内存,Windows 10 操作系统,编程环境 VC ++6.0,模式挖掘采用 joinlesss 方法。实验中数据集 1 采用合成数据,含类别特征 3 个,与分类任务相关空间特征 5 个,特征实例数共 2 000 个;数据集 2 为某地区电信服务运营商的服务类别空间数据,含类别特征 4 个(即 4 种服务类别),与分类任务相关空间特征 7 个,特征实例数共 3 000 个。其中,60% 的数据用于训练集,40% 的数据用于测试集。

1)分类准确率,结果见表 4。

表 4 几种算法分类准确率

比较项目	数据集 1	数据集 2
本文算法	87.2%	90.3%
文献[3]中的空间分类算法	83.1%	78.5%
文献[4]中的空间分类算法	84.5%	79.7%

实验结果表明,在特定的空间分类任务中,分类结果与待分类对象空间近邻的特征相关较大,与自身非空间属性相关较小,用文献[3]和文献[4]中传统的分类方法并不适用,分类准确率较低。本文提出的基于 co-location 模式的空间分类方法由于在训练阶段所得的分类规则均为兴趣度高的规则,分类准确率较高,是有效的空间分类方法。本实验在对数据集 1 和数据集 2 中的空间对象实现文献[3]和文献[4]的算法时,还合成了数据集 1,收集了数据集 2 中空间对象在分类时所需的数据,如空间对象的非空间属性、邻接关系、邻接对象的非空间属性等。

2)数据集 1 不同实例规模下的算法运行时间,结

果见表 5。

表 5　不同数据规模下算法运行时间

特征数	参与率	实例数	运行时间/ms
4	0.1	21	16
4	0.2	21	11
8	0.1	74	187
8	0.2	74	157
16	0.1	172	2 213
16	0.2	172	1 252
32	0.1	335	48 444
32	0.2	335	4 923

实验结果表明,随着特征数及实例数增加,算法所需的运行时间增长比较快,但也表明了本算法是高效的空间分类算法,可适用于空间数据库中的大数据集。

由于实例的近邻集中不含类别特征实例的完全独立连接与分类任务不相关,算法在数据预处理阶段,在实例的近邻集中去除不含类别特征实例的独立完全连接,在查找频繁模式的表实例时,搜索范围可得到有效的缩减。在分类规则集中,一部分规则在分类阶段的利用率不高,在搜索阶段却需要频繁搜索,增加了算法的开销,可以考虑提高模式参与率、统计规则使用率等方法减少分类规则。在分类之前还可以利用测试集对分类规则集进行划分,在测试集上测试分类准确率时挑选出频繁使用的规则,不频繁的规则可构成候选规则集,将测试集分成若干子集,重复挑选若干次,增加频繁规则集的兴趣度。在分类阶段,首先搜索频繁规则集,无规则适应的情况下再来搜索候选规则集即可。

4　结　语

在特定的空间分类任务中,类别与自身属性相关较小,与空间近邻对象的特征相关较大,用一些典型的空间分类方法进行分类并不适用,得到的分类准确率较低。基于 co-location 模式的空间分类算法由含类别特征的空间 co-location 模式导出与分类任务相关度比较高的分类规则,利用待分类对象空间近邻对象的特征对其分类。实验结果表明,本文提出的空间分类算法在特定的分类任务下是分类准确率较高的有效分类算法。但数据集增大时,算法的时间耗费增长较快,对算法进行有效剪枝,减少算法时间复杂度,提高分类准确率,是今后的努力方向。

参　考　文　献

[1] 张晶,毕佳佳,刘炉. 基于 mRMR 的多关系朴素贝叶斯分类[J]. 计算机应用与软件,2016,33(8):57－61.

[2] Fayyad R T, Muntz R. Mining Knowledge in Geographical Data[J]. IEEE Transaction on Knowledge and Data Engineering,2005,10:903－913.

[3] Ester M, Kriegel H P, Sander J. Spatial data mining: A database approach[C]//International Symposium on Advances in Spatial Databases. Springer-Verlag,1997:47－66.

[4] KoperSki K, Han J W, Stefanovic N. An efficient two-step method for classification of spatial data[J]. IEEE Transaction on Knowledge and Data Engineering,2008,14(5):1003－1016.

[5] Shekhar S, Schrater P R, Vatsavai R R, et al. Spatial contextual classification and prediction models for mining geospatial data[J]. IEEE Transactions on Multimedia,2002,4(2):174－188.

[6] Huang Y, Shekhar S, Xiong H. Discovering co-location patterns from spatial data sets: A general approach[J]. IEEE Transactions on Knowledge and Data Engineering,2004,16(12):1472－1485.

[7] Yoo J S, Shekhar S. A partial join approach for mining Co-location patterns[C]//Proceedings of the ACM International Symposium on Advances in Geographic Information System s (ACMGIS). Washington, USA,2004:241－249.

[8] Yoo J S, Shekhar S, Celik M. A join less approach for Co-location pattern mining: A summary of results[C]//Proceedings of the IEEE International Conference on Data Mining (ICDM). Houston, USA,2005:813－816.

[9] Wang Lizhen, Bao Yuzhen, Lu J, et al. A new join-less approach for co-location pattern mining[C]//Wu Qiang, He Xiangjian, Nguyen Q V. Proceedings of the IEEE 8th International Conference on Computer and Information Technology (CIT'08), Sydney, Australia, 2008. Piscataway, NJ, USA: IEEE,2008:197－202.

[10] Wang Lizhen, Bao Yuzhen, Lu Zhongyu. Efficient discovery of spatial co-location patterns using the iCPI-tree[J]. The Open Information Systems Journal,2009,3(1):69－80.

[11] Wang Lizhen, Zhou Lihua, Lu J. An order-clique-based approach for mining maximal co-locations[J]. Information Sciences,2009,179(19):3370－3382.

[12] 王丽珍,陈红梅. 空间模式挖掘理论与方法[M]. 北京:科学出版社,2014.

[13] 王丽珍,周丽华,陈红梅. 数据仓库与数据挖掘原理及应用[M]. 2 版. 北京:科学出版社,2009.

[14] 郭庆胜,魏智威,王勇,等. 特征分类与邻近图相结合的建筑物群空间分布特征提取方法[J]. 测绘学报,2017,46(5):631－638.