文章编号:1671-8860(2009)03-0331-04　　文献标志码:A

k-邻近空间关系下的空间同位模式挖掘算法

边馥苓[1]　万　幼[1]

(1　武汉大学空间信息与数字工程研究中心,武汉市珞喻路 129 号,430079)

摘　要:定义了一种基于 k-邻近对象的空间同位模式,探讨了基于 k-邻近空间关系的同位模式的特点及其与基于距离阈值的空间同位模式的区别与联系,并开发了 k-邻近对象同位模式挖掘算法(KNFCOM)。通过对真实数据的实验结果表明,KNFCOM 算法可有效地发现大型空间数据集中存在的各种空间同位模式。

关键词:k-邻近;空间同位模式;KNFCOM 算法;空间关联规则

中图法分类号:P208

空间同位模式是将关联规则泛化为空间索引的点集合数据集,把事务概念泛化以包括邻域集合,从地理空间中发现那些频繁的且紧密相邻的空间特征的集合[1]。空间同位模式问题的形式化描述如下。给定:① 一个包含 m 类空间特征类集合 $T=\{C_1, C_2, \cdots, C_m\}$;② 空间 S 内的 n 个空间特征对象 $O=\{o_0, o_1, \cdots, o_{n-1}\}$,每个空间特征对象用三元组标识 $o<$对象编号,特征类型,空间位置$>$,其中,对象编号$\in[0,n-1]$,特征类型$\in T$,空间位置$\in S$;③ 空间 S 内的一个特征邻近性规则 R。最终发现那些满足规则 R 且频繁出现的空间特征组,输出的空间同位模式可标识为多元组形式,一个二元同位模式的形式为 co-location(C_p, C_q),p、$q\in[1,m]$。

从现有的同位模式挖掘算法[2-7]来看,同位模式中对空间地物的点集合抽象显著地简化了空间关系的计算复杂度,并可有效地发现更为直接的空间模式。然而,这些算法在计算空间邻近关系时都是基于距离阈值的邻近度度量方式的,以发现频繁的邻近空间特征集合。这种度量方式存在以下局限性:① 为每个空间数据集选定距离阈值是困难的,且缺乏有效科学依据,通常只能靠经验或多次实验取得。② 空间数据集内的数据分布通常是不均匀的,因此使用给定的距离阈值来寻找邻近特征集合时,可能在数据分布稠密的区域产生太多的候选邻近特征集合,而在数据稀疏的区域产生很少的候选邻近特征集合,由此计算出的频繁邻近特征集合将无法准确地反映实际数据的空间同位情况。③ 对空间地物的点集合抽象可显著地简化空间关系的计算复杂度,然而对线状、面状地物进行点抽象之后,与之邻近的地物之间的距离可能变得较大,它们之间的邻近关系在基于距离的邻近度量方式中将消失。为此,本文定义了一种基于 k-邻近关系的空间同位模式,利用空间对象的 k-邻近对象集合度量其与其他对象的相似度,以空间对象的 k-邻近特征集合的成员支持度衡量空间对象与其邻近对象之间是否频繁,并发现其中存在的空间同位模式;同时开发了基于格网索引的 k-邻近特征同位模式挖掘算法(k nearest feature co-location mining, KNFCOM),可有效地发现大型空间数据集中存在的空间同位模式。

1　基于 k-邻近空间关系的同位模式

k-邻近关系被广泛地应用于数据挖掘领域的分类[8]和聚类[9]算法中。与数据挖掘领域常用的 k-邻近算法所不同的是,空间特征的相似度不是在任意对象之间进行计算,而是面向不同空间特征类之间的相似度度量。一个空间特征对象的 k-邻近空间特征对象是其与最邻近的 k 个其他类型的空间特征对象的集合。

在空间同位模式的形式化描述基础之上,使用欧几里德距离评估两个空间对象之间的距离

收稿日期:2009-01-16。

项目来源:国家自然科学基金资助项目(40601026)。

(标识为 dist(i, j)),给出定量描述的 k-邻近特征对象集合和频繁 k-邻近特征对象集合,引出 k-邻近关系下空间同位模式的量化标准。

定义1 特征对象的 k-邻近特征对象集合 k-NF(o)。一个特征对象 $o\in O$ 的 k-邻近特征对象集合 k-NF(o)是空间特征对象集 O 内与 o 最邻近的 k 个其他类型的空间特征对象,满足:① $|k\text{-NF}(o)|\leqslant k$;② $o\notin k\text{-NF}(o)$;③ 假设 o 的特征类型是 C,对象 p、$q\in k$-NF(o),那么 p、q 的特征类型不等于 C;④ 假设 p、q 分别是 o 的第 k 个、第 $k+1$ 个最邻近对象,那么 dist(o, q)$\geqslant$dist(o, p)。

定义2 特征类的 k-邻近特征类集合 k-NF(C)。一个特征类 $C\in T$ 的 k-邻近特征类集合 k-NF(C)是特征类 C 的所有特征对象 C_i 的 k-NF(C_i)的集合。

定义3 k-邻近特征类集合的成员支持度 Support(C, C')。它是特征类 C 的邻近特征类集合的某一成员 C' 在 C 的邻近特征类集合中出现的次数与邻近特征类集合 k-NF(C)大小的比值。成员支持度的值域是[0, 1]。

定义4 频繁 k-邻近特征类集合 Frequent(C,C')。给定 k-邻近特征类集合的成员支持度阈值 ε,若 C 的邻近特征类成员 C' 的支持度大于等于 ε,则认为此特征类成员 C' 在 C 的 k 邻近特征类集合中是频繁的。其形式化判定如下:若 $C\in k$-NF(C)且 Support(C,C') $\geqslant\varepsilon$,那么 Frequent(C,C')。

定义5 基于 k-邻近关系的空间同位模式 co-location$_k$($C\rightarrow C'$)。特征类 C 的频繁 k-邻近特征类集合的成员 C' 与 C 之间存在基于 k-邻近关系的空间同位模式。其形式化判定如下:若 $C'\in k$-NF(C)且 Frequent (C,C'),那么 co-location$_k$($C\rightarrow C'$)。

2 k-邻近空间同位模式的特点

2.1 非对称性

在基于距离阈值的同位模式发现方法中,当空间特征类 A 与 B 之间邻近关系的支持度 support(A, B)和 support(B, A)都满足最小支持度阈值,A 与 B 的邻近关系才是频繁的,存在 co-location(A, B)的同位模式。这种同位模式是对称的,即 co-location(A, B)$\Leftrightarrow$co-location(B,A)。然而,这种同位模式的计算很容易受 A、B 样本空间大小的影响,尤其是针对稀少空间特征的同位模式很难被发现。Huang[7] 提出了一种基于最大参与度的弱单调性算法发现稀少空间特征的同位模式,算法的本质是降低支持度阈值,以削弱算法的剪枝效果,但缺点是复杂度会随阈值的降低而呈几何级数增加。

根据定义3,k-邻近同位模式很好地考虑了不同空间特征类之间样本空间不一的事实,通常情况下,co-location$_k$(A,B)$\nLeftrightarrow$co-location$_k$(B,A),这种不对称性可用来解决针对稀少空间特征的同位模式发现的问题,并且效率较基于距离阈值的算法有显著提高。

2.2 包含性

基于 k-邻近空间同位模式实际上是基于距离阈值的空间同位模式的超集。从基于距离阈值发现空间同位模式的定义[2]来看,它发现的是空间数据集内属于不同特征类的空间特征对象频繁存在的紧密相邻的模式,其同位关系是双向的;而基于 k-邻近空间同位模式发现的是当某一类(或多类)空间特征对象出现时,其他特征类的对象频繁存在与其邻近位置的模式,其同位关系是单向的。若以相同的支持度阈值(ε)进行挖掘,从基于 k-邻近的空间同位模式中也可以推导出类似基于距离阈值的双向的空间同位模式:co-location$_k$($C\rightarrow C'$) $\wedge$ co-location$_k$($C'\rightarrow C$)$\Rightarrow$co-location(C,C')。

3 KNFCOM 算法实现与性能分析

3.1 算法实现描述

输入:① 空间特征类个数 m, m 是大于1的正整数;② 邻近对象个数 k,k 是大于0的正整数;③ k-邻近特征类的成员支持度阈值 ε,$\varepsilon\in(0,1)$;④ 空间特征对象集合 $O=\{A_1, A_2, \cdots, B_1, B_2, \cdots\}$。

输出:空间特征类之间存在的[2, $k+1$]元的空间同位模式。

算法实现:① 根据 m 值定义空间特征类集合 T 的大小,利用穷举法获取每个空间特征类 C 的 k-NF(C)的成员列表(共 $\sum_{i=1}^{k}C_{m-1}^{i}$ 个)。② 扫描空间特征对象集合 O,对于任一空间特征对象 o,$o\in C_i$,使用格网索引法计算 k-NF(o),o 的 k-NF(o)中应排除同属于特征类 C_i 的对象;按照所属的空间特征类,将在 k-NF(o)出现的各成员记录到 C_i 的 k-NF(C_i)中对应的位置,对重复出现的成员累加记数。③ 扫描空间特征类集合 T,对于每个空间特征类 C,扫描 k-NF(C),计算每个成员出现的支持度;若某一成员 C' 的支持度大于阈值 ε,那么

输出 co-location$_k$($C\rightarrow C'$)。④ 遍历所有的 k-邻近空间同位模式,对于一对一的 k-邻近空间同位模式 co-location$_k$($C\rightarrow C'$),若同时存在 co-location$_k$($C'\rightarrow C$),输出 co-location$_k$($C\leftrightarrow C'$);对于一对多的 k-邻近空间同位模式 co-location$_k$($C\rightarrow C'C''\cdots$),若对于 C'的任意子元素 C_i,都存在 co-location$_k$($C_i\rightarrow C$),输出 co-location$_k$($C\leftrightarrow C'C''\cdots$)。

3.2 算法性能分析

3.2.1 准确性

与基于距离阈值的邻近度度量方法相比,KNFCOM 算法使用基于 k-邻近对象的邻近度度量,具有更高的准确性。

1) 基于 k-邻近的空间关系度量方法限定的是邻近空间特征个数,基本不受空间数据分布的影响,较直接的距离度量更适应于空间分布不均匀的情况。

2) 通过设定不同的 k 值,和对 k-邻近特征对象集合的邻近对象进行特征类的排他性设置,KNFCOM 算法可以自定义地挖掘(k+1)元空间特征类之间的同位模式,且 k 值的设置仅对结果的同位模式中特征类的个数进行限定,不会对结果的准确性构成影响。而在基于距离阈值的同位模式发现算法中,为每个空间数据集选定距离阈值是困难的和缺乏有效科学依据的,并且距离阈值的设定对挖掘结果会有很大影响。

3) 在基于距离阈值的同位模式发现算法中,对空间地物的点集合抽象会忽略掉原本存在的点与线、点与面、线与面之间的邻近关系。文献[10]给出了一种针对扩展空间对象(线、面对象)的同位模式挖掘算法(extended co-location mining,EXCOM),此算法基于各种空间对象的最小外接矩形的缓冲区进行空间叠加分析,以度量空间对象之间的邻近关系,具有较高的计算复杂度,并且准确度受最小外接矩形大小和缓冲区面积大小的限制。在 KNFCOM 算法中,使用 k-邻近特征对象可以在一定程度上弱化因为线面的点集抽象造成的距离增大的效果。

此外,KNFCOM 算法对离群点也具有很好的判断和消解能力。通过在算法步骤②中使用格网索引法计算每个空间特征对象的 k-邻近对象时限定邻近格网的层数,可间接达到设定最大距离阈值的目的,即使在邻近格网层内未能找够 k 个邻近对象,也停止对其他格网邻近对象的搜索,以有效地防止那些在距离上显著离群的特征对象对挖掘结果的影响。

3.2.2 完整性

KNFCOM 算法所挖掘的空间同位模式是较为完整的,其完整性可从挖掘结果的广度和深度两方面进行分析。在广度上,KNFCOM 算法可有效地发现包括稀有空间特征类在内的各种空间特征类之间的同位模式,这是其优于基于距离阈值的同位挖掘算法的地方。在深度上,KNFCOM 算法能发现各种空间特征类直接单向的(不对称的)和双向的(对称的)不同类型的空间同位模式;而基于距离阈值的同位模式挖掘算法仅发现双向对称的同位模式。

3.2.3 复杂度

空间数据集的数据量通常不大,且空间特征的类型也不会很多,因此空间消耗不大,这里主要考虑算法的时间复杂度。

由于空间特征类集合 T 的元素个数 m 远小于空间特征对象集合 O 的元素个数 n,所以不涉及空间数据对象集合 O 的操作复杂度都可忽略。KNFCOM 算法的时间复杂度主要集中在步骤②扫描空间特征对象集合 O 计算 k-邻近特征对象集上,使用格网索引方法计算各空间特征对象的 k-邻近对象,其时间复杂度可控制在 $o(n)$内。

4 实验分析

在 Windows XP 操作系统下使用 Java 开发了 KNFCOM 算法,并在一台装有 奔腾 4-2.66 GHz、512 MB内存的台式机上进行了算法的实验分析。实验使用某港区地下各类管线的管线节点埋设情况作为空间数据集,分析各类管线节点间存在的 k-邻近空间同位模式。

4.1 地下管线节点间的空间同位模式发现

管线节点空间数据集共包含 5 类管线的 2 289个节点数据,见表 1。

表 1　管线节点空间数据集说明

Tab.1　Spatial Dataset of Pipe Nodes

管线节点类型	类型标识	对象个数
雨水管线节点	A	535
污水管线节点	B	110
给水管线节点	C	445
电力管线节点	D	861
电信管线节点	E	338

通过设定算法中不同的 k 值(2, 3, 4, 5)和支持度阈值 ε(60%, 70%),得到了各种管线节点的 k-邻近空间同位模式如表 2 所示(这里只列举 k=3、4,ε=0.6 时的结果)。

表2 管线节点空间数据集中的同位模式

Tab. 2 Co-location Patterns in Pipe Nodes Dataset

	$k=3$		$k=4$	
	co-location$_k$	支持度	co-location$_k$	支持度
支持度阈值0.6	$A \rightarrow D$	0.7	$A \rightarrow D$	0.7
	$B \rightarrow D$	0.7	$B \rightarrow A$	0.6
	$C \rightarrow A$	0.7	$B \rightarrow D$	0.8
	$C \rightarrow D$	0.7	$C \rightarrow A$	0.7
	$D \rightarrow A$	0.6	$C \rightarrow D$	0.8
	$E \rightarrow D$	0.9	$D \rightarrow A$	0.7
			$D \rightarrow E$	0.6
	$A \leftrightarrow D$	0.6	$E \rightarrow D$	0.9
			$A \leftrightarrow D$	0.7
			$D \leftrightarrow E$	0.6

由表2可知,在2 289个管线节点中,任一电信管线节点周围最邻近的三个管线节点中必有一个是电力管线节点的可能性是90%(单向的3-邻近空间同位模式);任一污水管线节点周围最邻近的三个管线节点中必有一个是电力管线节点的可能性是70%(稀少特征类的3-邻近空间同位模式);在所有雨水管线节点和所有电力管线节点中,它们互为4-邻近对象的可能性是70%(双向的4-邻近空间同位模式)。

4.2 参数 k 值对计算结果的影响

从表2的实验结果可知,随着 k 值的增大,所获得的空间同位模式也相应增加。图1显示,k 值对同位模式个数的影响基本是呈线性增长的关系。实际上,k 值取得过大,直接导致 k-邻近空间特征对象集合中的元素个数增多,因此所计算出的空间同位模式也增多。但由于 k-邻近空间特征对象集合中各对象间的邻近性因为 k 值的增大而被削弱了,这样会产生一些意义不大的模式,因此不赞成 k 的取值过大,最好取较空间特征类型个数略小的正整数。

4.3 数据集大小和 k 值对算法时间消耗的影响

对于算法的时间消耗,以 $m=5, k=2、3、4, \varepsilon=0.6$ 为参数,并选择三个大小不同的管线节点数据集,实验结果如图2所示。可见,随着数据集的增大,算法的时间消耗基本呈线性增长。而当 k 值在小于特征类型个数的一个范围内变化时,对算法的时间消耗影响很小。

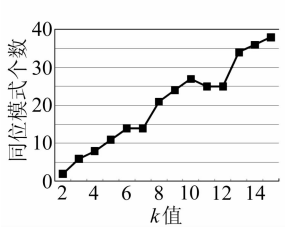

图1 空间同位模式个数与 k 值的关系图

Fig. 1 Influence of k Values on Patterns Generation

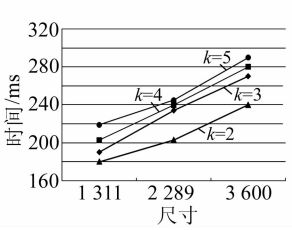

图2 数据集大小和 k 值对算法时间消耗的影响

Fig. 2 Influence of Data Size and k Values on Time Costs

5 结 语

本文定义了一种基于 k-邻近关系的空间同位模式,并开发了基于格网索引的 k-邻近特征同位模式挖掘算法(KNFCOM)。实验证明,KNFCOM算法可有效地发现大型空间数据集中存在的各种空间同位模式,对参数的设定和数据集的大小有很好的容忍度,并能解决在基于距离阈值的空间同位模式算法中难以解决的对稀少空间特征对象的同位模式的发现。以后的研究将着眼于扩展KNFCOM算法,以发现更多类型(多对一、多对多)的空间同位模式。此外,结合空间负关联关系发现空间异位模式,也是很有意义的研究内容。

参 考 文 献

[1] Shekhar S, Chawia S. 空间数据库[M]. 北京:机械工业出版社,2004

[2] Shekhar S, Huang Y. Co-location Rules Mining: A Summary of Results[C]. The 7th International Symposium on Spatio and Temporal Database (SSTD), New York, 2001

[3] Morimoto Y. Mining Frequent Neighboring Class Sets in Spatial Databases[C]. The 7th ACM SIGKDD International Conf on Knowledge Discovery and Data Mining, San Franciscc, California, 2001

[4] Huang Yan, Shashi S, Xiong Hui. Discovering Co-location Patterns from Spatial Datasets: A General Approach[J]. Transactions on Knowledge and Data Engineening, 2004,16(6):

[5] Yoo J, Shekhar S. A Partial Join Approach for Mining Co-location Patterns[C]. The 12nd Annual ACM International Workshop on Geographic Information Systems (ACM-GIS), Washington D C, USA, 2004

[6] Yoo J, Shekhar S, Celik M. A Join-less Approach for Co-location Pattern Mining: A Summary of Results[C]. The 5th IEEE International Conference on Data Mining(ICDM'05), Houston, USA, 2005

[7] Huang Yan, Pei Jian, Xiong Hui. Mining Co-Location Patterns with Rare Events from Spatial Data Sets[J]. GeoInformatica, 2006(10):239-260

(下转第338页)