计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201000186

星型高影响的空间 co-location 模式挖掘

马 董 李新源 陈红梅 肖 清
云南大学信息学院 昆明 650504
(md0301@mail.ynu.edu.cn)

摘 要 空间 co-location 模式是其实例在空间邻域内频繁并置出现的一组空间特征集。传统的空间 co-location 模式挖掘方法通常假设空间实例相互独立,并采用参与度作为模式有趣性的唯一度量指标,没有考虑不同特征或相同特征不同实例在空间邻域内所产生的影响差异,因此挖掘的结果往往缺乏相关性和可解释性。文中提出了一种星型高影响的空间 co-location 模式及挖掘方法,能够有效发现自身影响高且在邻域范围内也具有一定影响的空间 co-location 模式。首先,定义了度量模式影响的两个指标:模式影响参与度和模式影响占有度。其次,提出了挖掘星型高影响 co-location 模式的基础挖掘算法和剪枝策略。最后,通过在大量的真实和合成数据集上进行实验,分析了挖掘算法的效率和挖掘效果。实验结果表明,所提出的星型高影响 co-location 模式的度量方法和挖掘算法能够挖掘出较强相关性的 co-location 模式。

关键词:空间数据挖掘;空间 co-location 模式;星型影响;高影响模式
中图法分类号 TP391

Mining Spatial co-location Patterns with Star High Influence

MA Dong,LI Xin-yuan,CHEN Hong-mei and XIAO Qing
School of Information Science and Engineering,Yunnan University,Kunming 650504,China

Abstract The spatial co-location pattern is a group of spatial features whose instances are frequently collocated in the spatial neighborhood. Traditional spatial co-location pattern mining methods usually assume that the spatial instances are independent each other,and use participation index (PI) to measure the patterns. They don't consider the influence of different features or different instances of the same feature so that the mining results are often lack of relevance and interpretability. This paper proposes the spatial co-location pattern with star high influence which has influence in the neighborhood,and its mining method. Firstly,this paper defines two indicators to measure the influence of the pattern:influence participation index (IPI) and influence occupancy index (IOI). Secondly,a basic algorithm and pruning strategies for mining co-location patterns with star high influence are proposed. Finally,the experimental results on real and synthetic data sets show that the proposed method can discover the strong relevant co-location patterns.

Keywords Spatial data mining,Spatial co-location pattern,Star influence,High influence pattern

1 引言

随着互联网信息技术的飞速发展,全球定位系统、社交媒体、移动设备的普遍使用,数据的产生和收集越来越自动化,导致空间数据存储量暴增。如何从海量的空间数据中有效地挖掘出隐藏的、有价值的知识信息和具有预测性的规律,从而给人类的生产生活提供正确的指导和科学的决策,成为了亟待解决的问题,空间数据挖掘的出现很好地满足了需要。空间 co-location 模式作为空间数据挖掘的重要子领域,主要目的是发现频繁共现的空间特征组,该组特征的实例频繁并置出现[1]。例如,火车站附近往往存在宾馆,松茸往往生长在长苞冷杉附近。挖掘空间 co-location 模式能够从空间数据中获取有价值的、具有指导性的空间并置关系,且能应用于道路交通[2]、城市计算[3-4]、环境监测[5]等诸多领域。

传统的空间 co-location 模式挖掘方法大多将空间实例视为具有相同地位的独立个体,通过团实例模型计算参与度并将其作为评价模式有趣性的唯一指标。然而,在实际生活中,同属学校类型的一所大学和一所小学的社会影响是不一样的;同属超市类型的一个大型超市和一个小卖部受人们的关注度也是不一样的,只关注模式的频繁性,而不考虑不同特征或相同特征不同实例的影响差异性,可能会导致挖掘结果缺乏实用性。另外,传统挖掘方法也未考虑模式在邻域内的影

到稿日期:2020-10-30 返修日期:2021-03-25 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61662086,61966036);云南省创新团队项目(2018HC019)
This work was supported by the National Natural Science Foundation of China(61662086,61966036) and Project of Innovative Research Team of Yunnan Province(2018HC019).
通信作者:陈红梅(hmchen@ynu.edu.cn)

响，即模式外的特征和实例对模式的影响，可能会导致挖掘结果缺乏真正的相关性，在城市规划中，由于城市 POI(Point Of Interest)的高密度分布和强自相关性，如果不考虑模式外的特征和实例对模式的影响，可能会将一些实际上并无相关性的模式报告给用户。

传统团实例模型要求模式中的实例两两邻近构成团，忽略了空间特征间重要的非团的空间关系。而星型实例模型放松了团约束，仅需模式中实例与中心特征实例邻近，从而可以发现空间特征间更丰富的空间关系研究特征。基于以上考虑，本文以星型实例模型为基础，研究带有影响值的空间数据，提出了一种星型高影响的空间 co-location 模式及挖掘方法。本文主要贡献总结如下：

(1)基于星型实例模型，提出了一种星型高影响的空间 co-location 模式，并定义了两个度量模式影响性的指标：模式影响参与度和模式影响占有度。

(2)提出了一种挖掘星型高影响 co-location 模式的基础算法，并设计了基于特征最小影响参与率的剪枝策略。

(3)通过在真实和合成数据集上的实验，分析本文算法的效率和挖掘结果，并验证了该模式的相关性和可解释性。

2　相关工作

自 Shekhar 等[6]提出 co-location 模式及基于完全连接的 join-based 算法以来，许多研究者针对 co-location 模式的概念、算法及应用进行了深入研究，co-location 模式大致可分为：效率型模式挖掘、数据驱动型模式挖掘、目标驱动型模式挖掘。

(1)效率型 co-location 模式挖掘

join-based 算法在生成候选模式及表实例时，大量的连接操作可能会导致算法的时间开销大，而 join-less 算法[7]将实例间的邻近关系物化为星型邻居关系，用查询操作代替连接操作来获取表实例，避免在生成表实例阶段的大量连接操作。为了提高 co-location 模式挖掘的数据规模和效率，一些研究人员将挖掘算法搭建于并行平台之上。比如，文献[8]将 co-location 模式挖掘算法运行于 MapReduce 平台上，从而提高了 co-location 模式的挖掘效率。文献[9]在并行环境下采用有序团的方法挖掘 co-location 模式，其能够直接得出模式的表实例，不需要判断是否满足团关系，大大提升了算法的效率。

(2)数据驱动型 co-location 模式挖掘

针对不确定数据，文献[10]通过概率密度函数来描述位置不确定的空间实例，将 join-based 算法扩展为 ujoin-based 算法，挖掘位置不确定空间数据中的 co-location 模式。针对模糊数据，文献[11]提出了模糊参与率和模糊参与度来挖掘模糊空间 co-location 模式。文献[12]将密度峰值聚类算法和模糊理论相结合来实现实例对簇的模糊划分，并采用模糊团代替传统团来挖掘 co-location 模式。针对时空数据，文献[13]将实例存在的时间区间作为约束条件，重新定义了邻近关系，挖掘带有时间约束的模式。文献[14]将时空事件之间的时间间隔的影响引入时空 co-location 模式的兴趣度量中，提出了一种加权的滑动窗口模型来挖掘时空 co-location 模式。

(3)目标驱动型 co-location 模式挖掘

为了使 co-location 模式更加简洁实用，一些研究以挖掘闭频繁和极大频繁 co-location 模式为目标。文献[15]考虑到空间特征间的交互，提出了一种超参与度的闭模式，实现了对模式的精简无损表达。文献[16]将所有二阶频繁 co-location 模式生成一个无向稀疏图，然后从图中找出所有的极大团，从而挖掘更加简易的极大频繁模式。文献[17]引入了一种极大候选模式树，提出了一种挖掘前 k 个最长的极大 co-location 模式的方法，由于该方法无须生成所有的候选模式，因此减少和节省了挖掘算法所需的时间和空间。为了识别出传统 co-location 模式挖掘方法可能会遗漏的信息，文献[18]通过计算特征在模式与子模式间的参与率变化，提出了一种含有主导特征的 co-location 模式。文献[19-20]引入星型参与实例代替团实例，提出了亚频繁 co-location 模式的概念及挖掘方法。针对领域知识的 co-location 模式挖掘，部分研究者引入效用概念。文献[21]将效用引入到 co-location 模式挖掘中，考虑不同特征具有不同的效用，相同特征不同实例的效用相同，进而提出了高效用 co-location 模式及基本挖掘框架。

3　传统 co-location 模式挖掘的相关概念

出现在地理空间中不同类型的对象称为空间特征，通常用 $F=\{f_1,f_2,\cdots,f_n\}$ 表示 n 个空间特征的集合。出现在具体地理位置上的对象称为空间实例，常用 $S=\{S_1\cup S_2\cup\cdots\cup S_n\}$ 表示 n 个空间特征的实例集，其中 $S_i(1\leqslant i\leqslant n)$ 表示特征 f_i 对应的实例集。如果两个实例 $i_i,i_j\in S$ 的欧几里得距离小于或等于用户指定距离阈值 d，那么称这两个实例满足空间邻近关系 R，即 $R(i_i,i_j)\Leftrightarrow dis(i_i,i_j)\leqslant d$。给定一个 k 阶空间 co-location 模式 $c=\{f_1,f_2,\cdots,f_k\}(c\subseteq F,k=|c|)$，实例集 $I=\{i_1,i_2,\cdots,i_k\}(I\subseteq S)$ 满足形成一个团，即 $\{R(i_i,i_j)|1\leqslant i\leqslant k,1\leqslant j\leqslant k\}$，如果 I 包含 c 的所有实例且不存在 I 的任何子集包含 c 的实例，那么称 I 为 c 的一个行实例，记为 $row\_instance(c)$，c 的所有行实例构成 c 的表实例，记为 $table\_instance(c)$。如图 1 所示，图中有 A,B,C,D 4 个空间特征，特征 A 有 3 个实例，特征 B 有 4 个实例，特征 C 有 3 个实例，特征 D 有 2 个实例，满足邻近关系的实例用实线连接，模式 $\{A,B,C\}$ 的表实例 $table\_instance(\{A,B,C\})=\{\{A.1,B.1,C.1\},\{A.2,B.2,C.2\}\}$。

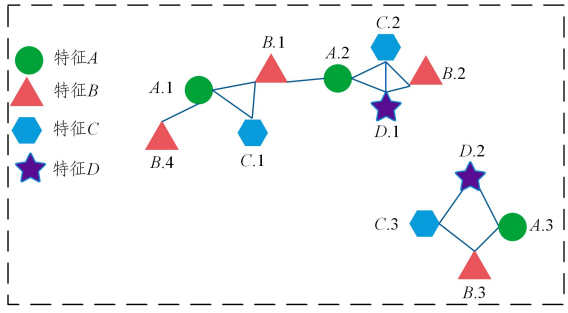


图 1　空间特征及其实例示例

Fig. 1　Example of spatial features and instances

传统 co-location 模式挖掘方法通常采用参与率(Participation Rate，PR)度量特征的重要性，采用参与度(Participa-

tion Index,PI)度量模式的有趣性。特征 f_i 在模式 c 中的参与率定义为 f_i 的不重复实例个数与其实例总个数的比率,表示为:

$$PR(c,f_i)=\frac{|\pi_{f_i}(table\_instance(c))|}{|S_{f_i}|} \tag{1}$$

其中,π 是关系投影操作,S_{f_i} 是特征 f_i 的所有实例集合。

模式 c 的参与度定义为 c 中所有特征参与率的最小值,表示为 $PI(c)=\min_{i=1}^{k}\{PR(c,f_i)\}$。当 $PI(c)$ 大于或等于用户指定的最小参与度阈值 min_prev 时,则称模式 c 为频繁 co-location 模式。在图 1 中,模式 $\{A,B,C\}$ 的参与度 $PI(\{A,B,C\})=\min\{2/3,1/2,2/3\}=1/2$,假如参与度阈值 $min\_prev=0.4$,那么模式 $\{A,B,C\}$ 为一个频繁 co-location 模式。

4 星型高影响 co-location 模式的相关定义及度量指标

本节将给出带影响值的空间实例、特征影响参与率、特征影响权重等定义,并给出两个度量模式影响的指标:模式影响参与度和模式影响占用度。影响参与度用于度量模式内部的影响情况,影响占有度用于度量模式在邻域内(模式外)的影响情况。

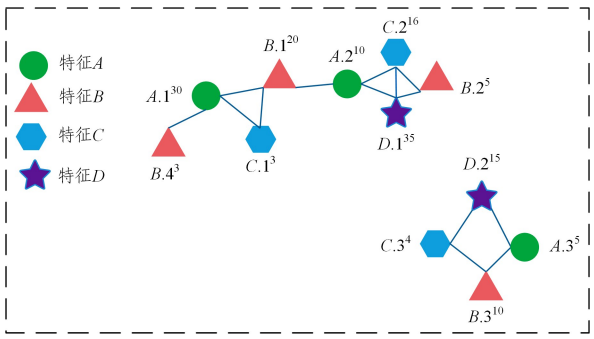


图 2 带影响值的空间实例示例

Fig. 2 An example of spatial instances with influence value

定义 1(带影响值的空间实例) 带有影响 v 的空间特征 f_i 的第 j 个实例记为 $f_i.j^v$,或者将实例 $f_i.j$ 的影响记为 $e(f_i.j)=v$。例如,图 2 中带影响的实例 $A.1^{30}$ 的影响值为 $e(A.1)=30$。

定义 2(特征影响参与率) 给定一个 k 阶的 co-location 模式 $c=\{f_1,f_2,\cdots,f_k\}$,$f_i\in c(1\leqslant i\leqslant k)$ 在模式 c 中的影响参与率定义为:以 f_i 为中心的星型参与实例中不重复的 f_i 实例的影响值之和与 f_i 的所有实例的影响值总和的比率。其表示为:

$$FIR(c,f_i)=\frac{\sum_{f_i.j\in SPIns(c,f_i)}e(f_i.j)}{E(f_i)} \tag{2}$$

其中,$E(f_i)$ 表示特征 f_i 的所有实例的影响值总和。

例如,在图 2 中,模式 $\{A,B,C\}$ 中各个特征的影响参与率分别为:

$$FIR(\{A,B,C\},A)=\frac{30+10}{30+10+5}=0.89$$

$$FIR(\{A,B,C\},B)=\frac{20+10}{20+5+10+3}=0.79$$

$$FIR(\{A,B,C\},C)=\frac{3+16}{3+16+4}=0.83$$

特征影响参与率既考虑了特征参与到模式的实例数量,也考虑了不同实例的影响,以此综合判断特征在模式中的影响。然而特征间的影响也是不一样的,那么我们将通过特征影响权重来反映特征间的影响。

引理 1(特征影响参与率的反单调性) 特征在模式中的影响参与率随模式阶数的增大而单调递减。

证明:假设 $f_i.j\in SPIns(c,f_i)$ 表示实例 $f_i.j$ 是特征 f_i 在模式 c 中的一个星型参与实例。

那么,模式 c 的子模式 $c'\subset c$,也一定满足 $f_i.j\in SPIns(c',f_i)$,所以,$\sum_{f_i.j\in SPIns(c,f_i)}e(f_i.j)\leqslant\sum_{f_i.j\in SPIns(c',f_i)}e(f_i.j)$,

$$\frac{\sum_{f_i.j\in SPIns(c,f_i)}e(f_i.j)}{E(f_i)}\leqslant\frac{\sum_{f_i.j\in SPIns(c',f_i)}e(f_i.j)}{E(f_i)}$$

$$FIR(c,f_i)\leqslant FIR(c',f_i)$$

因此特征影响参与率是单调递减的。

定义 3(特征影响权重) 给定一个 k 阶的 co-location 模式 $c=\{f_1,f_2,\cdots,f_k\}$,$f_i\in c(1\leqslant i\leqslant k)$ 在模式 c 中的影响权重定义为:以 f_i 为中心的星型参与实例中不重复的 f_i 实例的影响值之和与模式 c 中所有特征的星型参与实例中不重复实例的影响值总和的比率。其表示为:

$$\lambda(c,f_i)=\frac{\sum_{f_i.j\in SPIns(c,f_i)}e(f_i.j)}{E(c)} \tag{3}$$

其中,$E(c)$ 表示模式 c 中所有特征的星型参与实例中不重复实例的影响值总和。

例如,在图 2 中,模式模式 $\{A,B,C\}$ 中的各个特征的影响权重为:

$$\lambda(\{A,B,C\},A)=\frac{30+10}{89}=0.45$$

$$\lambda(\{A,B,C\},B)=\frac{20+10}{89}=0.34$$

$$\lambda(\{A,B,C\},C)=\frac{3+16}{89}=0.21$$

模式中各个特征的影响权重之和为 1。

定义 4(模式影响参与度) 给定一个 k 阶的 co-location 模式 $c=\{f_1,f_2,\cdots,f_k\}$,模式 c 的影响参与度定义为:模式 c 中所有特征的影响参与率与特征影响权重乘积的总和。其表示为:

$$IPI(c)=\sum_{i=1}^{k}FIR(c,f_i)\times\lambda(c,f_i) \tag{4}$$

例如,在图 2 中,模式 $\{A,B,C\}$ 的影响参与度 $IPI(\{A,B,C\})=0.89\times0.45+0.79\times0.34+0.83\times0.21=0.84$。

定义 5(星型参与实例邻居) 给定一个 k 阶的 co-location 模式 $c=\{f_1,f_2,\cdots,f_k\}$,实例 $f_i.j$ 是特征 $f_i\in c(1\leqslant i\leqslant k)$ 在模式 c 中的一个星型参与实例,那么实例 $f_i.j$ 的星型参与实例邻居定义为:以实例 $f_i.j$ 为中心,模式 c 中其余特征的实例与 $f_i.j$ 有邻近关系的所有实例的集合,记为 $SPIns\_neighbor(c,f_i.j)$。

例如,在图 2 中,特征 A 的实例 $A.1$ 的星型参与实例邻居 $SPIns\_neighbor(\{A,B,C\},A.1)=\{B.1,B.4,C.1\}$。表 1 列出了模式 $\{A,B,C\}$ 中的各个特征的星型参与实例邻居。

表 1　模式$\{A,B,C\}$的星型参与实例邻居

Table 1　Star-participation-instances' neighbors of $\{A,B,C\}$

特征类型	星型参与实例	星型参与实例邻居
A	$A.1$	$\{B.1,B.4,C.1\}$
	$A.2$	$\{B.1,C.2\}$
B	$B.1$	$\{A.1,A.2,C.1\}$
	$B.3$	$\{A.3,C.3\}$
C	$C.1$	$\{A.1,B.1\}$
	$C.2$	$\{A.2,B.2\}$

定义 6(共影响实例)　给定一个 k 阶的 co-location 模式 $c=\{f_1,f_2,\cdots,f_k\}$，实例 $f_i.j$ 是特征 $f_i\in c(1\leqslant i\leqslant k)$在模式 c 中的一个星型参与实例，实例 $f_i.j$ 的共影响实例定义为：与 $f_i.j$ 的星型参与实例邻居都有邻近关系的在模式 c 以外的实例。f_i的所有星型参与实例的共影响实例构成了 f_i的共影响实例，记为 $ce\_ins(c,f_i)$。

例如，在图 2 中，模式$\{A,B,C\}$中各个特征的共影响特征实例为 $ce\_ins(\{A,B,C\},B)=\{D.2^{15}\}$，$ce\_ins(\{A,B,C\},C)=\{D.1^{35}\}$。

特征 B 的共影响实例 $ce\_ins(\{A,B,C\},B)=\{D.2^{15}\}$，说明特征 D 的一个影响值为 15 的实例减弱了模式中特征 B 的实例对特征 A 和 C 的实例的影响。特征 A 没有共影响特征实例，说明特征 A 的实例在与特征 B 和 C 的实例共现时，不受其他实例的影响，即特征 A 真正影响了模式$\{A,B,C\}$的形成。

定义 7(影响占有率)　给定一个 k 阶的 co-location 模式 $c=\{f_1,f_2,\cdots,f_k\}$，特征 $f_i\in c(1\leqslant i\leqslant k)$的影响占有率定义为：以 f_i为中心的不重复星型参与实例的影响值总和，占其与 f_i的共影响实例的影响值总和之和的比率。表示为：

$$IOR(c,f_i)=\frac{\sum\limits_{f_i.j\in SPIns(c,f_i)}e(f_i.j)}{\{\sum\limits_{f_i.j\in SPIns(c,f_i)}e(f_i.j)\}+\{\sum\limits_{f_n.m\in ce\_ins(c,f_i)}e(f_n.m)\}}\tag{5}$$

例如，在图 2 中，模式$\{A,B,C\}$中各个特征的影响占有率为：

$$IOR(\{A,B,C\},A)=\frac{20+10}{20+10}=1$$

$$IOR(\{A,B,C\},B)=\frac{20+10}{20+10+15}=0.67$$

$$IOR(\{A,B,C\},C)=\frac{3+16}{3+16+35}=0.35$$

虽然特征 A 与特征 B 在模式$\{A,B,C\}$中的影响参与率相差不大，但是特征 A 在模式$\{A,B,C\}$中的星型参与实例不受其他实例的影响，即特征 A 在模式$\{A,B,C\}$中不存在共影响实例，说明特征 A 和 B 对模式$\{A,B,C\}$的影响存在差异。那么影响占有率通过计算特征参与到模式中的实例与模式外的实例(共影响实例)之间的影响情况，能够进一步衡量特征在邻域内对模式的影响。

定义 8(影响占有度)　给定一个 k 阶的 co-location 模式 $c=\{f_1,f_2,\cdots,f_k\}$，c 的影响占有度定义为 c 中所有特征的影响占有率的最小值。其表示为：

$$IOI(c)=\min_{i=1}^{k}(\{IOR(c,f_i)\})\tag{6}$$

例如，在图 2 中，模式$\{A,B,C\}$的影响占有度为 $IOI(\{A,B,C\})=\min(1,0.67,0.35)=0.35$。

模式的影响参与度能度量模式内部的影响，而模式的影响占有度能度量模式在邻域内的影响。我们将通过这两个度量指标挖掘模式本身影响高且在邻域内影响也高的星型高影响模式。

定义 9(星型高影响 co-location 模式)　给定一个 k 阶的 co-location 模式 c、最小影响参与度阈值 min_ipi、最小影响占有度阈值 min_ioi，若满足如下条件：1) $IPI(c)\geqslant min\_ipi$；2) $IOI(c)\geqslant min\_ioi$。那么模式 c 就为一个星型高影响 co-location 模式。

例如，在图 2 中，假设给定阈值 $min\_ipi=0.5$，$min\_ioi=0.3$。模式$\{A,B,C\}$的影响参与度 $IPI(\{A,B,C\})=0.84$，影响占有度 $IOI(\{A,B,C\})=0.34$，那么$\{A,B,C\}$是一个星型高影响 co-location 模式。

引理 2(星型高影响 co-location 模式不满足先验原理)　模式 c 不是星型高影响 co-location 模式，其超模式不一定不是星型高影响 co-location 模式。即 IPI 和 IOI 不会随着阶的增大而逐渐减小。

证明：(举反例)假设 $min\_ipi=0.5$，$min\_ioi=0.3$。模式$\{A,C\}$的影响参与度 $IPI(\{A,C\})=0.83$，影响占有度 $IOI(\{A,C\})=0.25$，所以$\{A,C\}$不是星型高影响 co-location 模式，但是它的高阶模式$\{A,B,C\}$是星型高影响模式，因为 $IPI(\{A,B,C\})=0.84$，$IOI(\{A,B,C\})=0.35$。可见，星型高影响 co-location 模式不满足先验原理。

定理 1　给定一个 co-location 模式 $c=\{f_1,f_2,\cdots,f_k\}$，如果 c 中的每个特征的影响参与率都小于 min_ipi，即 $\forall f_i\in c|\{FIR(c,f_i)<min\_ipi\}$，那么 c 一定不是星型高影响 co-location 模式。

证明：如果 $FIR(c,f_i)<min\_ipi$，那么：

$$FIR(c,f_i)\times\lambda(c,f_i)<min\_ipi\times\lambda(c,f_i)$$

$$\begin{aligned}IPI(c)&=\sum_{i=1}^{k}FIR(c,f_i)\times\lambda(c,f_i)\\&=FIR(c,f_1)\times\lambda(c,f_1)+\cdots+FIR(c,f_k)\times\lambda(c,f_k)\\&<min\_ipi\times\lambda(c,f_1)+\cdots+min\_ipi\times\lambda(c,f_k)\\&=min\_ipi\times(\lambda(c,f_1)+\cdots+\lambda(c,f_k))=min\_ipi\end{aligned}$$

因此当模式 c 中的每个特征的影响参与率都小于 min_ipi 时，模式的参与影响度也一定小于 min_ipi，那么模式 c 一定不是星型高影响 co-location 模式。

例如，在图 2 中，假设 $min\_ipi=0.8$，模式$\{B,D\}$中的特征影响参与率为：$FIR(\{B,D\},B)=0.13$，$FIR(\{B,D\},D)=0.7$，所以模式$\{B,D\}$不是星型高影响 co-location 模式。

5　挖掘算法及剪枝策略

由于星型高影响 co-location 模式不满足向下闭合性质，因此本文不能采用与传统 co-location 模式挖掘方法类似的先验知识进行有效剪枝。基于星型高影响 co-location 模式的性质，本节首先给出一个基于“候选-测试”方法的星型高影响 co-location 模式的基础挖掘算法；其次提出一种基于最小影响参与率的剪枝策略，改进算法的挖掘效率。

5.1 基础算法

星型高影响 co-location 模式的基础挖掘算法可分为以下 3 个步骤。

步骤 1 将输入的带有影响的空间数据进行物化处理，得到星型邻居集；

步骤 2 生成 $k(k\geqslant 2)$ 阶候选星型高影响 co-location 模式；

步骤 3 计算候选模式的影响参与度和影响占有度，与用户指定的阈值进行比较，筛选得出星型高影响 co-location 模式集。

星型高影响 co-location 模式的基础挖掘算法如算法 1 所示。

算法 1 Basic-ASHICM 算法

输入：空间特征集 F、带影响值的实例集 I、邻近距离阈值 d、最小影响参与度阈值 min_ipi、最小影响占有度阈值 min_ioi

输出：星型高影响 co-location 模式集 SHI_set

变量：co-location 模式阶数 k、k 阶候选星型高影响 co-location 模式 Con_SHI_k

```
BEGIN
1. 生成星型邻居集 SN;
2. k=2,生成 2 阶候选模式 Con_SHI_k;
3. WHILE(Con_SHI_k≠NULL)
4.   FOR EACH c∈Con_SHI_k
5.     计算特征的影响参与率 FIR(c,f_i)及影响权重 λ(c,f_i);
6.     计算模式 c 的影响参与度 IPI(c);
7.     IF IPI(c)≥min_ipi
8.       计算特征的的影响占有率 IOR(c,f_i);
9.       计算模式 c 的影响占有度 IOI(c);
10.      IF IOI(c)≥min_ioi
11.        将模式 c 加入到高影响模式集 SHI_set;
12.    生成 k+1 阶候选模式 Con_SHI_{k+1};
13.  k=k+1;
14. 输出星型高影响 co-location 模式集 SHI_set。
END
```

5.2 剪枝策略

模式的度量指标为影响参与度和影响占有度，都不满足反单调性原理，因此无法采用传统的剪枝策略。为了提高挖掘效率，我们根据特征影响参与率所满足的反单调性，设计了基于最小影响参与率的剪枝算法。

定义 10（相关模式） 给定一个 k 阶 co-location 模式 c，如果某个 k 阶模式与模式 c 有 $k-1$ 个特征相同，那么称该模式为模式 c 的一个相关模式，所有的相关模式构成了模式 c 的相关模式集，记为 $CLp(c)$。

定义 11（相关特征） 给定一个 co-location 模式 c，c 的相关模式 c_r 与 c 不同的特征称为模式 c 的相关特征，即相关特征 $f_i\in c_r, c_r\in CLp(c)$ 且 $f_i\notin c$，将所有相关特征称为 c 的相关特征集，记为 $CLf(c)$。

例如，在图 2 中，模式$\{A,B\}$的相关模式集 $CLp(\{A,B\})=\{\{A,C\},\{A,D\},\{B,C\},\{B,D\}\}$。相应地，模式$\{A,B\}$的相关特征集 $CLf(\{A,B\})=\{C,D\}$。

定理 2 给定一个 k 阶模式 $c=\{f_1,f_2,\cdots,f_k\}$，如果 c 满足以下两个条件，那么 c 及其所有超模式 $c_s\supset c$ 都不可能是星型高影响模式，可以无须计算直接剪枝。

(1) $\forall f_i\in c|\{FIR(c,f_i)<min\_ipi\}$；

(2) $\max\{\min_{f_j\in CLf(c)}[FIR(c_r,f_j),c_r\in CLp(c)]\}<min\_ipi$。

证明：假设模式 c 的超模式 $c_s=c\cup f$，f 是 c 的相关特征在相关模式 $CLp(c)$ 中的最小影响参与率中值最大的特征，且 f 的影响参与率小于 min_ipi，则有：

$$\begin{aligned}IPI(c_s)&=\sum_{f_j\in c_s}\{FIR(c_s,f_j)\times\lambda(c_s,f_j)\}\\&=\{\sum_{f_i\in c\wedge f_i\in c_s}(FIR(c_s,f_i)\times\lambda(c_s,f_i))\}+FIR(c_s,f)\times\lambda(c_s,f)\end{aligned}$$

依据引理 1，影响参与率随模式阶数的增大而单调递减，且 $\forall f_i\in c|\{FIR(c,f_i)<min\_ipi\}$。所以有：

$$\begin{aligned}IPI(c_s)&\leqslant\sum_{f_i\in c\wedge f_i\in c_s}\{FIR(c,f_i)\times\lambda(c_s,f_i)\}+FIR(c_s,f)\times\lambda(c_s,f)\\&<min\_ipi\times(1-\lambda(c_s,f))+FIR(c_s,f)\times\lambda(c_s,f)\\&<min\_ipi\times(1-\lambda(c_s,f))+min\_ipi\times\lambda(c_s,f)\\&=min\_ipi\end{aligned}$$

综上所得，当模式 c 满足上述两个条件时，模式 c 及其所有超模式都不是星型高影响 co-location 模式。

例如，在图 2 中，假设影响参与度阈值 $min\_ipi=0.8$，模式$\{B,D\}$的相关模式集 $CLp(\{B,D\})=\{\{A,B\},\{B,C\},\{C,D\},\{A,D\}\}$，相关特征集 $CLf(\{B,D\})=\{A,C\}$。计算得到 $\max\{\min_A(1,0.33),\min_C(1,0.77)\}=0.77<0.8$，因此模式$\{B,D\}$及其所有超模式都不是星型高影响 co-location 模式。

挖掘星型高影响 co-location 模式时，可通过影响参与度和影响占有度两个指标逐项度量，当影响参与度大于或等于给定阈值后，再进行影响占有度的相关计算及阈值比较。基于定理 2，我们利用影响参与度这一指标进行剪枝，设计了基于最小影响参与率的剪枝算法（Min Feature Influence Ratio Algorithm，MFIRA），其具体描述如算法 2 所示。

算法 2 MFIRA 算法

输入：空间特征集 F、带影响值的实例集 I、邻近距离阈值 d、最小影响参与度阈值 min_ipi、最小影响占有度阈值 min_ioi

输出：星型高影响 co-location 模式集 SHI_set

变量：co-location 模式阶数 k、k 阶候选星型高影响 co-location 模式 Con_SHI_k

```
BEGIN
1. 生成星型邻居集 SN;
2. k=2,生成所有 2 阶候选模式 Con_SHI_k;
3. WHILE(Con_SHI_k≠ NULL)
4.   FOR EACH c ∈Con_SHI_k
5.     计算特征的影响参与率 FIR(c,f_i)及影响权重 λ(c,f_i);
6.     计算模式 c 的影响参与度 IPI(c);
7.     IF IPI(c)<min_ipi
8.       使用定理 2 检查模式 c 的超模式是否可以剪枝，如果满足剪
         枝条件，将 c 加入到剪枝模式集 P_P 中;
9.     ELSE
10.      计算特征的的影响占有率 IOR(c,f_i)
11.      计算模式 c 的影响占有度 IOI(c);
12.      IF IOI(c)≥min_ioi
13.        将 c 加入到高影响模式集 SHI_set 中;
14. 生成 k+1 阶的候选模式 Con_SHI_{k+1}，并基于剪枝模式集 P_P，使
```

用定理 2 对候选模式 Con_SHI_{k+1} 进行剪枝;

15. k=k+1;

16. 输出星型高影响 co-location 模式集 SHI_set。

END

6　实验结果与分析

本节将在合成数据和真实数据集上进行实验分析,评估本文提出的星型高影响 co-location 模式的基础挖掘(Basic Algorithm,BA)算法和剪枝 MFIRA 算法。实验的主要目标是:1)在合成数据集上,分析 BA 算法与 MFIRA 算法的运行效率受不同实验参数的影响;2)在真实数据集上,评估两个度量指标对挖掘算法 MFIRA 挖掘结果的影响,并分析模式实例。

6.1　实验设置

实验数据:实验采用 2 个真实数据集和 4 个合成数据集来评价算法性能和验证挖掘效果,数据集的统计信息如表 2 所列。其中,Beijing-POI 来自北京市 POI 数据,包含 16 种 POI 类型(空间特征),23 025 个具体 POI(空间实例),分布形状如图 3 所示。Plantdata 数据是来自"三江并流"区域的珍稀植物数据集,包含了 31 种植物类型(空间特征),336 株具体植物(空间实例),分布形状呈带状如图 4 所示。合成数据集采用泊松分布函数生成。

表 2　数据集信息

Table 2　Dataset information

数据集	特征数	实例数	分布范围	影响值
Beijing-POI	16	23 025	22 000×14 000	[1,100]
Plantdata	31	336	80 000×130 000	[1,100]
合成数据集 1	20	20 000	1 000×1 000	[1,100]
合成数据集 2	20	20 000	2 000×2 000	[1,100]
合成数据集 3	20	40 000	2 000×2 000	[1,100]
合成数据集 4	40	80 000	2 000×2 000	[1,100]

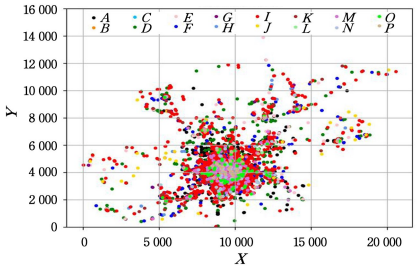


图 3　北京市 POI 数据集分布图

Fig. 3　Distribution of Beijing POI' dataset

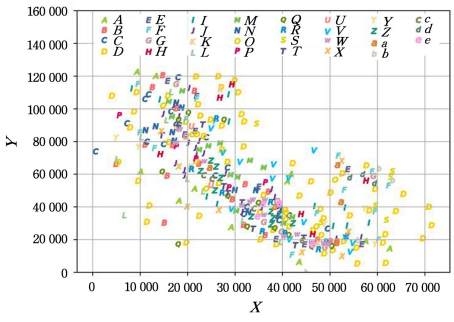


图 4　"三江并流"区域珍稀植物数据集分布图

Fig. 4　Distribution of "Three Parallel Rivers" rare plants' dataset

运行环境:本文中所有算法都采用 python 语言实现,并运行于具有英特尔酷睿 i7 CPU、8 GB 内存、500 GB 存储硬盘、Windows 10 及 pycharm2017 的 PC 机上。

实验参数:算法在各个数据集上的实验参数的默认设置如表 3 所列。

表 3　实验参数默认值

Table 3　Default value of experimental parameters

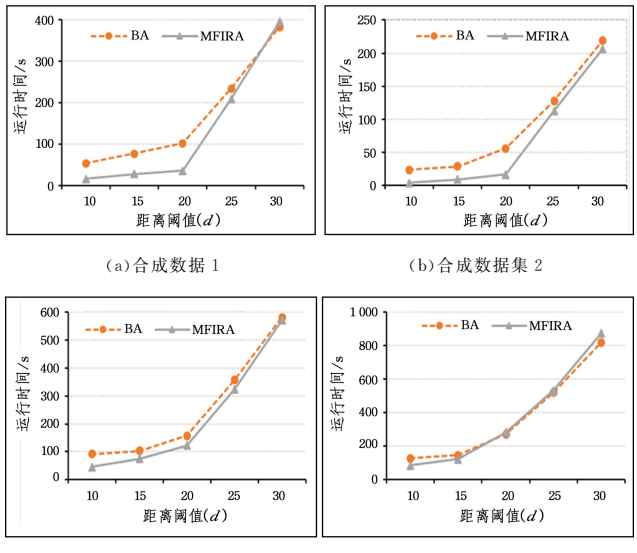
数据集	距离阈值	影响参与阈值	影响占有度
Beijing-POI	50	0.5	0.4
Plantdata	5 000	0.5	0.4
合成数据集 1	20	0.5	0.4
合成数据集 2	20	0.5	0.4
合成数据集 3	20	0.5	0.4
合成数据集 4	20	0.5	0.4

6.2　算法效率分析

我们在合成数据集 1—4 上分析不同参数变化下,星型高影响 co-location 模式的基础挖掘算法 BA 与剪枝算法 MFIRA 的性能。主要目的是通过合成数据集 1 和 2 分析在不同的数据分布密度下,算法受实验参数的影响;通过合成数据集 3 和 4 分析在不同数量的特征和实例下,算法受实验参数的影响。

6.2.1　距离阈值对算法运行效率的影响

针对距离阈值 d,分别在 4 个合成数据集上设置 5 个不同的阈值(10,15,20,25,30)来观察距离阈值对算法运行时间的影响,其余参数取表 3 中的默认值,图 5 给出了在 4 个合成数据集上算法运行时间随不同距离阈值的变化情况。对于所有数据集,随着距离阈值的增大,算法运行时间逐渐增多,剪枝效果也逐渐变差甚至运行时间多于基础算法。这是因为随着距离阈值增大,星型邻居随之增多,算法也就更耗时;特征的影响参与率也随之增大,基于最小影响参与率的剪枝算法可识别的剪枝模式变少,剪枝算法在判断可剪枝条件上花费了更多的时间。



(a)合成数据 1　(b)合成数据集 2

(c)合成数据集 3　(d)合成数据集 4

图 5　不同距离阈值下算法运行时间

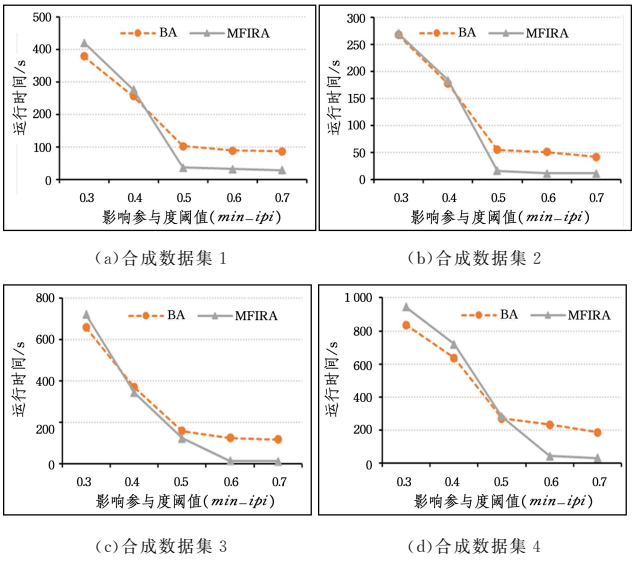
Fig. 5　Time cost with different d

从图 5(a)、图 5(b)观察得出,合成数据集 1 的运行时间

比合成数据集 2 的运行时间长，但其剪枝效果却没有合成数据集 2 好。这是因为合成数据集 1 和 2 的特征与实例数相同，但是分布范围不相同，合成数据集 1 的数据密度更高，实例之间更易满足邻近关系。从图 5(c)、图 5(d)观察得出，随着数据集规模的增大，算法运行时间也逐渐增长，剪枝算法在合成数据集 4 上效果较差，剪枝算法在数据集较大时剪枝效率不高。这是因为特征数和实例数较多时，剪枝算法在判断剪枝条件时，需要计算更多的相关特征在相关模式中的影响参与率值。

6.2.2 影响参与度阈值对算法运行效率的影响

针对影响参与度阈值 min_ipi，分别在 4 个合成数据集上设置 5 个不同的阈值(0.3，0.4，0.5，0.6，0.7)来观察影响参与度阈值对算法运行时间的影响，其余参数取表 3 中的默认值，图 6 给出了在 4 个合成数据集上算法运行时间随不同影响参与度阈值的变化情况。对于所有数据集，随着影响参与度阈值的增大，算法运行时间逐渐增长，剪枝效率随之提高。这是因为剪枝算法识别剪枝条件是比较特征影响参与率与影响参与度阈值，而影响参与度阈值增大，满足剪枝的模式就越多。从图 6(a)、图 6(b)同样可以看出合成数据集 2 的运行效率优于合成数据集 1，体现了数据的密集度对算法的影响。从图 6(c)、图 6(d)同样可以看出合成数据集 4 的运行效率低于合成数据集 3，体现了数据规模对算法的影响。



(a)合成数据集 1 (b)合成数据集 2

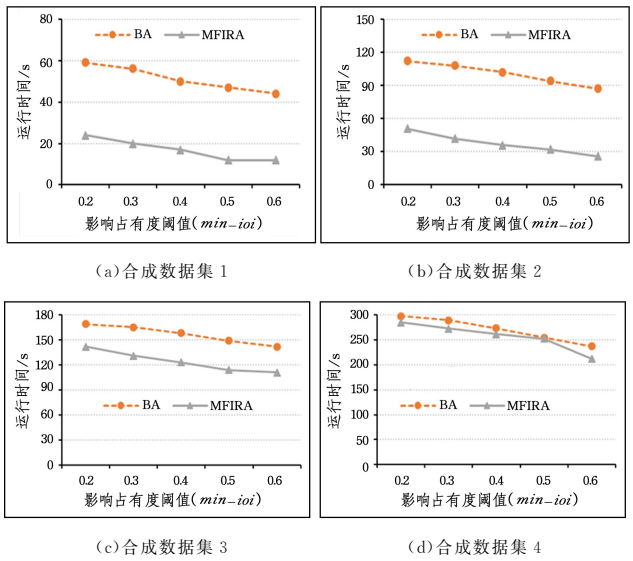
(c)合成数据集 3 (d)合成数据集 4

图 6 不同影响参与度阈值下算法运行时间

Fig. 6 Time cost with different min_ipi

6.2.3 影响占有度阈值对算法运行效率的影响

针对影响占有度阈值 min_ioi，分别在 4 个合成数据集上设置 5 个不同的阈值(0.2，0.3，0.4，0.5，0.6)来观察影响占有度阈值对算法运行时间的影响，其余参数取表 3 中的默认值，图 7 给出了在 4 个合成数据集上算法运行时间随不同影响占有度阈值的变化情况。对于所有数据集，随着影响占有度阈值的增大，两种算法的运行时间均小幅减少。这是因为算法先判断邻近关系，再检验影响参与度与相关阈值的关系，最后才检验影响占有度与相关阈值的关系，所以当距离阈值和影响参与度阈值确定后，需要剪枝的模式也被确定，影响占有度对算法效率的影响就不明显。合成数据集 1—3 的剪枝效果比合成数据集 4 要好。这是因为数据集较大时，特征影响参与率受距离阈值与 min_ioi 的影响较大，满足特征影响参与率小于 min_ioi 的特征较少，所以算法的剪枝效率不高。



(a)合成数据集 1 (b)合成数据集 2

(c)合成数据集 3 (d)合成数据集 4

图 7 不同影响占有度阈值下算法运行时间

Fig. 7 Time cost with different min_ioi

6.3 挖掘结果分析

通过设置两个度量指标的不同数值来分析挖掘算法在真实数据集上的结果。

6.3.1 影响参与度对挖掘结果的影响

图 8 给出了不同影响参与度 min_ipi 下星型高影响 co-location 模式的挖掘数量，其余参数取表 3 中的默认值。从图 8 可以看出，随着影响参与度的增大，模式数量逐渐减少。值得注意的是，当 $min\_ipi=0.5$ 时，POI 数据的变化幅度比珍稀植物数据大，这是因为 POI 数据分布为中心密集、四周稀疏，当影响度达到一个临界值时有大量的模式不满足星型高影响 co-location 模式的要求，而珍稀植物数据分布较均匀，模式的参与度分布也较均匀。

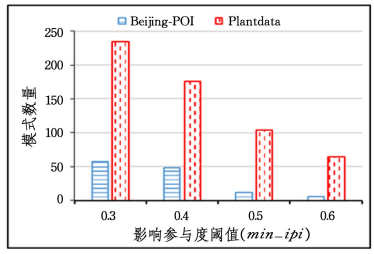


图 8 不同影响参与度阈值下的模式数量

Fig. 8 Number of patterns with different min_ipi

6.3.2 影响占有度对挖掘结果的影响

图 9 给出了不同影响占有度 min_ioi 下星型高影响 co-location 模式的挖掘数量，其余参数取表 3 中的默认值。从图 9 可以看出，随着影响占有度的增大，模式数量逐渐减少。则 POI 数据的模式数量明显少于珍稀植物数据，这是因为影响占有度考虑了邻域内不相关特征对模式的影响，而 POI 数据分布较密集，数据本身自相关性强，模式外不相关特征较多，所以模式的影响占有度较低。

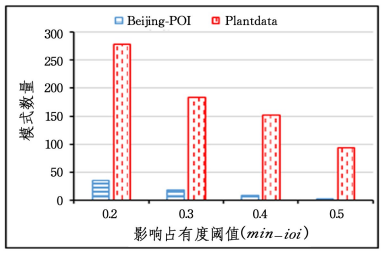


图 9　不同影响占有度阈值下的模式数量

Fig. 9　Number of patterns with different *min_ioi*

6.4　模式实例分析

如表 4 所列,我们选取了算法在真实数据集上的部分模式实例进行分析。

表 4　真实数据集上的挖掘示例

Table 4　Mining example on real dataset

星型高影响 co-location 模式	PEI	EOI
{中餐馆(*A*),宾馆(*D*),招待所(*E*)}	0.74	0.56
{中餐馆(*A*),花园(*F*),服装店(*M*)}	0.60	0.52
{云南榧木(*C*),兰类(*D*),云南红豆杉(*J*),贡山三尖杉(*M*)}	0.65	0.49
{虫草(*Q*),梭砂贝母(*R*),天女花(*T*)}	0.57	0.51

图 10 和图 11 分别给出了星型高影响模式{中餐馆(*A*),宾馆(*D*),招待所(*E*)}、{中餐馆(*A*),花园(*F*),服装店(*F*)}的实例分布。

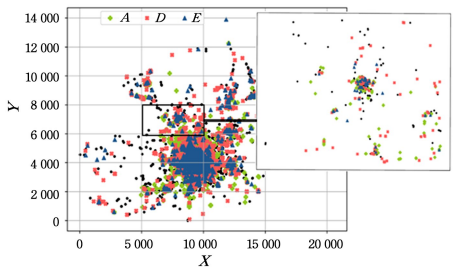


图 10　模式{*A*,*D*,*E*}的实例分布

Fig. 10　Distribution of {*A*,*D*,*E*}'s instances

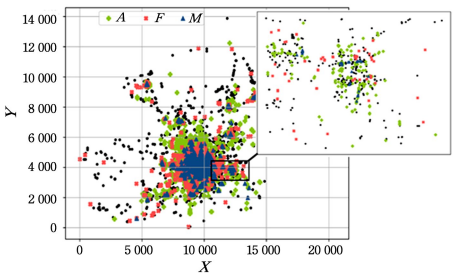


图 11　模式{*A*,*F*,*M*}的实例分布

Fig. 11　Distribution of {*A*,*F*,*M*}'s instances

从图 10 可以看出,模式{*A*,*D*,*E*}的实例在商业中心和郊区都有分布,这是因为{中餐馆,宾馆,招待所}这一模式在生活中是为了满足人们的基本生活需求,所以无论是市中心还是郊区都普遍存在。从右上角的放大图可以看出,模式的实例在邻域内分布集较密集,且在邻域内其他实例较少,说明了该模式在邻域内具有较高的影响。

从图 11 可以看出,模式{*A*,*F*,*M*}的实例在商业中心分布较多,而在郊区分布较少,这是因为模式中的花园和服装店一般规划于市中心,而郊区由于交通不便利和人流量少,一般很少存在花园和服装店。如右上角放大的图所示,模式的实例分布同样较密集,模式的实例在邻域内所占比例较大,说明了模式{*A*,*F*,*M*}在邻域内的影响也较高。

图 12 和图 13 分别给出了星型高影响模式{云南榧木(C),兰类(D),云南红豆杉(J),贡山三尖杉(M)}和{虫草(Q),梭砂贝母(R),天女花(T)}的实例分布。

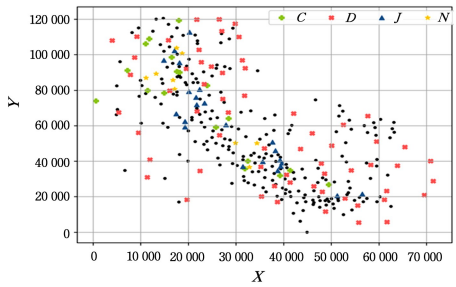


图 12　模式{*C*,*D*,*J*,*M*}的实例分布

Fig. 12　Distribution of {*C*,*D*,*J*,*M*}'s instances

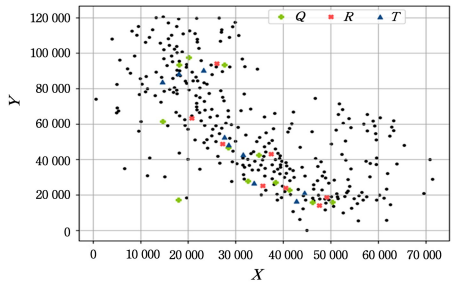


图 13　模式{*Q*,*R*,*T*}的实例分布

Fig. 13　Distribution of {*Q*,*R*,*T*}'s instances

从图 12 中可以看出,模式在邻域内实例占比较大,说明模式{*C*,*D*,*J*,*M*}在邻域内具有较大的影响性。从图 13 中可以看出,模式{*Q*,*R*,*T*}在邻域内的实例占比不高,这是因为植物数据集中各个特征的实例数较少,实例所带的影响值对挖掘结果的影响较大,而本文是通过随机方法给每个实例生成影响值,所以就会出现类型{*Q*,*R*,*T*}的模式,模式的实例在邻域内密度不大,但是所带的影响值较大,因此该模式在邻域内也具有较大的影响。

结束语　针对传统的空间 co-location 模式挖掘方法未考虑空间特征或者实例的差异性,导致挖掘结果往往缺乏针对性的问题,本文充分考虑了不同特征、同一特征不同实例的差异,既分析了模式内的特征对模式的影响,也分析了模式在邻域内的影响,进而提出了一种星型高影响 co-location 模式及度量方法。由于无法满足向下闭合性原理,本文设计了基础挖掘算法,并基于特征影响参与率满足的反单调性,设计了合理的剪枝策略。通过在真实数据集和合成数据集上的实验,证明了所提算法的有效性,并分析了真实数据集上的模式实例。在未来的研究工作中,可以在本文的基础之上,结合实例的实际影响,设计更加合理的影响值;将挖掘算法搭建在并行平台上,以增大实验数据的规模;设计更高效的剪枝策略,提高算法的挖掘效率。

参 考 文 献

[1] WANG L Z, CHEN H M. Spatial Pattern Mining Theory and Methods[M]. Beijing: Science Press, 2014: 2-4.

[2] AN S, YANG H, WANG J, et al. Mining urban recurrent congestion evolution patterns from GPS equipped vehicle mobility data[J]. Information Sciences, 2016, 373: 515-526.

[3] WU C F, CAI L, LI J, et al. Frequent Pattern Mining of Residents' Travel Based on Multi-source Location Data[J]. Computer Science, 2021, 48(7): 155-163.

[4] SUN T X, ZHAO Y L, LIAN Z W, et al. Mobility Pattern Mining for People Flow Based on Spatio-Temporal Data[J]. Computer Science, 2020, 47(10): 91-96.

[5] AKBARI M, SAMADAZDEGAN F, WEIBEL R. A generic regional spatio-temporal co-occurrence pattern mining model: a case study for air pollution[J]. Journal of Geographical Systems, 2015, 17(3): 249-274.

[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 join-less approach for mining spatial colocation patterns[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(10): 1323-1337.

[8] YOO J S, BOULWARE D, KIMMEY D. A parallel spatial co-location mining algorithm based on MapReduce[C]// 2014 IEEE International Congress on Big Data. IEEE, 2014: 25-31.

[9] YANG P, WANG L, WANG X. A parallel spatial co-location pattern mining approach based on ordered clique growth[C]// International Conference on Database Systems for Advanced Applications. Cham: Springer, 2018: 734-742.

[10] LU Y, WANG L Z, ZHANG X F. Mining frequent co-location patterns from uncertain data[J]. Journal of Frontiers of Computer Science and Technology, 2009, 3(6): 656-664.

[11] OUYANG Z P, WANG L Z, CHEN H M. Mining spatial co-location patterns for fuzzy objects[J]. Chinese Journal of Computers, 2011, 34(10): 1947-1955.

[12] FANG Y, WANG L, HU T. Spatial co-location pattern mining based on density peaks clustering and fuzzy theory[C]// Proceedings of the 2018 Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data, LNCS 10988. Cham: Springer, 2018: 298-305.

[13] ZENG X, YANG J. Co-location patterns mining with time constraint[J]. Computer Science, 2016, 43(2): 293-296.

[14] QIAN F, YIN L, HE Q, et al. Mining spatio-temporal co-location patterns with weighted sliding window[C]// 2009 IEEE International Conference on Intelligent Computing and Intelligent Systems. IEEE, 2009, 3: 181-185.

[15] WANG L, BAO X, CHEN H, et al. Effective lossless condensed representation and discovery of spatial co-location patterns[J]. Information Sciences, 2018, 436: 197-213.

[16] YAO X, PENG L, YANG L, et al. A fast space-saving algorithm for maximal co-location pattern mining[J]. Expert Systems with Applications, 2016, 63: 310-323.

[17] BAO X, WANG L, ZHAO J. Mining top-k-size maximal co-location patterns[C]// 2016 International Conference on Computer, Information and Telecommunication Systems (CITS). IEEE, 2016: 1-6.

[18] FANG Y, WANG L, WANG X, et al. Mining co-location patterns with dominant features[C]// International Conference on Web Information Systems Engineering. Cham: Springer, 2017: 183-198.

[19] WANG L, BAO X, ZHOU L, et al. Maximal sub-prevalent co-location patterns and efficient mining algorithms[C]// Proceedings of the 2017 International Conference on Web Information Systems Engineering, LNCS 10569. Cham: Springer, 2017: 199-214.

[20] WANG L, BAO X, ZHOU L, et al. Mining maximal sub-prevalent co-location patterns[J]. World Wide Web, 2019, 22(5): 1971-1997.

[21] YANG S S, WANG L Z, LU J L, et al. Primary Exploration for Mining Spatial High Utility Co-location Patterns[J]. Journal of Chinese Mini-Micro Computer Systems, 2014, 35(10): 2302-2307.

MA Dong, born in 1992, master. His main research interests include spatial data mining and so on.

CHEN Hong-mei, born in 1976, Ph. D, associate professor. Her research interests include database and spatial data mining.

(责任编辑:柯颖)