计算机科学
COMPUTER SCIENCE

空间co-location模式的主导特征挖掘

熊开放，陈红梅，王丽珍，肖清

引用本文

熊开放，陈红梅，王丽珍，肖清.空间co-location模式的主导特征挖掘[J]．计算机科学, 2022, 49(11A): 211000126-7.

XIONG Kai-fang, CHEN Hong-mei, WANG Li-zhen, XIAO Qing. Mining Spatial co-location Pattern with Dominant Feature [J]. Computer Science, 2022, 49(11A): 211000126-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211000126

空间 co-location 模式的主导特征挖掘

熊开放　陈红梅　王丽珍　肖　清
云南大学信息学院　昆明 650000

摘　要　空间 co-location 模式是空间特征的子集,它们的实例在邻域内频繁并置出现。传统 co-location 模式不区分模式中特征的重要性,忽略了特征间的主导关系。主导特征 co-location 模式考虑模式中特征的不平等性,分析特征间的主导关系,具有重要的应用意义。然而,现有主导特征模式挖掘没有从特征实例分布的角度综合考虑一个特征主导其他特征的可能倾向和影响强度,使得挖掘的主导特征及模式没有较好地反映特征间的主导关系。首先分析 co-location 模式中特征实例的空间分布,提出模式主导度,用以度量模式中某个特征主导其他特征的可能倾向;提出主导影响度,用以度量模式中某个特征主导其他特征的影响强度;基于这两个新度量,提出 co-location 模式的主导特征挖掘。然后通过优化新度量的计算,提出有效的主导特征 co-location 模式挖掘算法。在真实数据集和合成数据集上开展大量实验,验证了所提方法能够有效地识别 co-location 模式中的主导特征,所提算法能够高效地挖掘主导特征及模式。

关键词:空间数据挖掘;空间 co-location 模式;主导特征;主导特征模式
中图法分类号　TP301

Mining Spatial co-location Pattern with Dominant Feature

XIONG Kai-fang,CHEN Hong-mei,WANG Li-zhen and XIAO Qing
School of Information Science and Engineering,Yunnan University,Kunming 650000,China

Abstract　A spatial co-location pattern is a subset of spatial features whose instances frequently locate together in the neighborhood. Traditional co-location pattern does not distinguish the importance of features in the pattern,and ignores the dominant relationship among features. The co-location pattern with dominant feature considers the inequality of features in the pattern,and analyzes the dominant relationship among features,which can be used in many applications. However,the existing methods for mining co-location pattern with dominant feature do not comprehensively consider the possible tendency and influence intensity of one feature dominating other features from the perspective of features' instances distribution,so that the dominant relationship among features is not properly revealed. This paper first analyzes the spatial distribution of features' instances in a co-location pattern,proposes the pattern dominance index to measure the possible tendency of a feature dominating other features in a pattern,and proposes the dominant influence index to measure the influence intensity of the dominance tendency. Based on the two new measures,the dominant feature mining of co-location pattern is proposed. Then an efficient algorithm for mining co-location pattern with dominant feature is proposed by optimizing the calculation of new measures. A large number of experiments on real data sets and synthetic data sets verify that the proposed method can effectively identify the dominant feature in a co-location pattern,and it can efficiently mine co-location patterns with dominant feature.

Keywords　Spatial data mining,Spatial co-location pattern,Dominant feature,Pattern with dominant feature

1　引言

空间 co-location 模式是其实例在空间中频繁关联的空间特征的子集,其思想源于地理学第一律"任何事物都与其他事物相关,但相近的事物关联更紧密"[1]。Co-location 模式揭示了空间特征间的频繁并置关系,例如,火车站附近往往有旅馆;西尼罗河病毒往往出现在蚊子泛滥、饲养家禽的区域[2]。由于 co-location 模式在环境保护[3]、城市计算[4]、公共交通[5]等领域中的广泛应用,因此 co-location 模式挖掘吸引了研究者的广泛关注,形成了空间数据挖掘的重要研究方向。传统 co-location 模式挖掘采用特征参与度作为模式的有趣性度量,该度量考虑了特征实例参与到模式实例中的比例,揭示了空间特征间的频繁并置关系,但忽略了空间特征间的主导关系。然而,在现实中,co-location 模式中的空间特征具有不同的重要性,某些特征的出现主导了其他特征的出现,从而使得它们频繁并置出现。例如,模式{医院,药店,花店}揭示了医院、药店和花店频繁并置出现,但不能揭示它们之间的主导关系,即医院主导了花店和药店的出现,花店和药店依赖于医院而存在。进一步挖掘 co-location 模式中的主导特征,揭示模式中空间特征间的主导关系,有助于深入分析模式中特征间的共生关系、排斥关系、因果关系等重要关系,为空间数据分析及决策提供支持。

然而,现有 co-location 模式主导特征挖掘虽然考虑了模式中特征的不平等性,但是没有从特征实例分布的角度综合

基金项目:国家自然科学基金(61662086,61762090,61966036)
This work was supported by the National Natural Science Foundation of China(61662086,61762090,61966036).
通信作者:陈红梅(hmchen@ynu.edu.cn)

考虑一个特征主导其他特征的可能倾向和影响强度。直观地说，在 co-location 模式中，一个特征实例的邻域有更多其他特征的实例，这个特征主导其他特征的倾向就越大，而其他特征实例与这个特征实例的距离越近，这个特征主导其他特征的强度就越大。例如，在如图 1 所示的空间数据集中，有医院、药店和花店 3 个空间特征，分别记为 A，B，C，医院的实例集为{$A.1$}，药店的实例集为{$B.1$，$B.2$，$B.3$，$B.4$}，花店的实例集为{$C.1$，$C.2$，$C.3$，$C.4$}。医院实例 $A.1$ 的邻域内有许多药店和花店的实例{$B.1$，$B.2$，$B.3$，$C.1$，$C.2$，$C.3$}，反映医院的存在主导了药店和花店的出现。此外，$A.1$ 对这些实例的影响强度随着距离的增加而减弱。基于此，本文分析 co-location 模式中特征实例的空间分布，综合特征间主导关系的可能倾向和影响强度，提出两个新的度量以有效地识别模式中的主导特征。

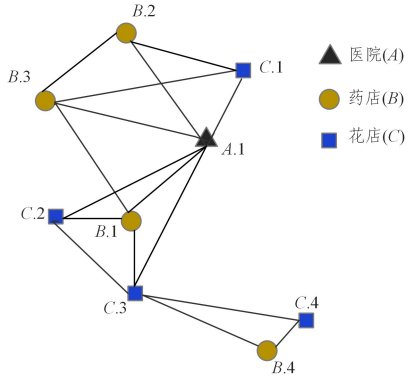


图 1 一个空间实例集

Fig. 1 Space instance set

本文主要工作归纳如下：

(1)提出模式主导度，用以度量模式中某个特征主导其他特征的可能倾向；提出主导影响度，用以度量模式中某个特征主导其他特征的影响强度；基于这两个新度量，提出 co-location 模式的主导特征挖掘。

(2)通过优化新度量的计算，提出有效的主导特征 co-location 模式挖掘算法。

(3)在真实数据集和合成数据集上开展了大量实验，验证了所提方法能够有效地识别 co-location 模式中的主导特征，所提算法能够高效地挖掘主导特征及模式。

2 相关工作

Huang[6] 等最先提出了空间 co-location 模式挖掘的基本概念参与率和参与度以及基本算法 join-based。参与度是参与率的最小值，与度量空间特征对之间相互作用的统计方法 cross-K 函数密切相关。虽然 join-based 算法利用参与度的反单调性缩减模式搜索空间来提高模式挖掘效率，但是 join-based 算法的大量连接操作仍导致算法整体效率不高。为了提高 join-based 算法效率，研究者提出了一系列优化算法。Partial-join[7] 算法和 join-less[8] 算法以减少生成表实例的连接操作为目标而提出，CPI-tree[9] 算法、ICPI-tree[10] 算法和 Order-Clique-Base[11] 算法以优化候选搜索空间和表实例存储空间为目标而提出。此外，研究者根据数据的不同特性，扩展了传统 co-location 模式。针对数据的模糊性，文献[12]提出了模糊参与率和模糊参与度概念，文献[13-14]基于模糊集理论与聚类算法，提出模糊模式挖掘算法。针对带效用的特征或实例，文献[15]提出了高效用模式挖掘方法。进一步考虑数据的时变性，文献[16]提出了时空数据集上高效用模式的增量挖掘方法。针对特征实例的分布特性，Huang[17] 等最先提出用最大参与率作为度量以挖掘稀有特征模式。由于最大参与率没有兼顾模式的频繁性，文献[18]提出了基于最小加权参与率的稀有特征模式挖掘。

在考虑空间 co-location 模式中特征间的主导关系方面，文献[19]分析了空间特征及实例的耦合关系，通过特征参与到模式及其子模式的实例变化，度量特征间的影响差异，进而挖掘主导特征及模式；文献[20]首先识别含有关键特征的显著 co-location 模式，然后在显著 co-location 模式中挖掘含有多个主导特征的模式。文献[19-20]是从基于团实例模型的 co-location 模式中挖掘主导特征及模式。与文献[19-20]不同，文献[21]从基于星型实例模型的亚频繁 co-location 模式中挖掘主导特征及模式。

本文研究基于团实例模型的 co-location 模式中的主导特征挖掘，但与文献[19-20]不同，本文分析 co-location 模式中特征实例的空间分布，并综合特征间主导关系的可能倾向和影响强度，以度量模式中的主导特征。

3 基本概念及问题定义

3.1 基本概念

给定空间特征集合 $F=\{f_1,f_2,\cdots,f_n\}$，空间实例集合 $S=S_1\cup S_2\cup\cdots\cup S_n$，其中 $S_i(1\leqslant i\leqslant n)$ 是特征 f_i 的实例集合，以及距离阈值 d。如果任意两个实例 i_i，$i_j\in S$ 的距离小于等于 d，则称实例 i_i，i_j 满足空间邻近关系 R。对于 k 阶空间 co-location 模式 $c=\{f_1,f_2,\cdots,f_k\}(c\subseteq F)$，如果实例集 $I=\{i_1,i_2,\cdots,i_k\}(I\subseteq S)$ 满足 I 中实例的特征集包含 c 中所有特征且 I 形成团，即 I 中任意两个实例满足空间邻近关系 R，则称 I 为 c 的一个行实例，c 的所有行实例构成 c 的表实例，记为 $T(c)$。特征 f_i 在模式 c 中的参与率定义为 $PR(f_i,c)=\dfrac{|\pi_{f_i}(T(c))|}{|T(\{f_i\})|}$，其中 π 是关系投影操作。模式 c 的参与度定义为 $PI(c)=\min\limits_{i=1}^{k}\{PR(f_i,c)\}$。空间 co-location 模式挖掘就是从空间特征及实例集中挖掘参与度 PI 大于等于给定的最小参与度阈值 min_prev 的所有频繁模式。在本文中，如果不加特别说明，所指模式均为频繁模式。

3.2 问题定义

为了识别 co-location 模式中的主导特征，本文分析模式中特征实例的空间分布，并综合特征间主导关系的可能倾向和影响强度，提出两个新的度量：模式主导度和主导影响度。

定义 1(特征主导率) 给定一个 k 阶模式 $c=\{f_1,f_2,\cdots,f_i,\cdots,f_k\}$，特征 f_i 在模式 c 中的主导率 $FDP(f_i,c)$ 定义为：

$$FDP(f_i,c)=\frac{\min\{PR(f_j,c)\mid f_j\in c,f_j\neq f_i\}}{|\pi_{f_i}T(c)|}\tag{1}$$

其中，$PR(f_i,c)$ 是特征 f_i 在模式 c 中的参与率，$T(c)$ 是模式 c 的表实例。特征主导率 $FDP(f_i,c)$ 表示在模式 c 表实例中，特征 f_i 的每个实例吸引其他特征实例的最小比率。模式 c 中某个特征的主导率越大，这个特征主导其他特征的可能倾向越大。

定义 2(模式主导度) 给定一个 k 阶模式 $c=\{f_1,f_2,\cdots,f_i,\cdots,f_k\}$，模式 c 的主导度 $PDI(c)$ 定义为模式 c 中

特征的最小主导率，即：

$$PDI(c)=\min\{FDP(f_i,c)\mid f_i\in c\} \tag{2}$$

模式 c 的主导度表示模式 c 中特征的主导倾向的下界。

定义 3（候选主导特征模式和候选主导特征） 给定一个 k 阶模式 $c=\{f_1,f_2,\cdots,f_k\}$ 及模式主导度阈值 min_pdi，如果 $PDI(c)\geqslant min\_pdi$，则模式 c 称为候选主导特征模式，模式中特征主导率最大的特征构成候选主导特征集合 G，即 $G=\{\arg\max_{f_i\in c} FDP(f_i,c)\}$。

例 1 图 1 所示的数据集中，对于模式 $c=\{A,B,C\}$，$PDI(c)=\min\{FDP(A,c),FDP(B,c),FDP(C,c)\}=\min\{0.75,0.25,0.25\}=0.25$。如果模式主导度阈值为 0.2，则模式 c 就是候选主导特征模式，且候选主导特征集为 $\{A\}$。

定义 4（分布均值） 设模式 $c=\{f_1,f_2,\cdots,f_k\}$ 是候选主导特征模式，其中的候选主导特征集为 Gc，模式的表实例为 $T(c)$，对于候选主导特征 $f_i\in G$ 的实例 $i_i\in\pi_{f_i}T(c)$，模式中其他特征 $f_j\in c$ 的实例相对于 i_i 的分布均值 $DV(i_i,f_j,c)$ 定义为：

$$DV(i_i,f_j,c)=\frac{\sum_{i_j\in\pi_{f_j}cl(i_i)}D(i_i,i_j)}{|\pi_{f_j}cl(i_i)|} \tag{3}$$

其中，$cl(i_i)$ 为 $T(c)$ 中包含 i_i 的行实例集合，$D(i_i,i_j)$ 是 i_i 与 i_j 之间的距离。

定义 5（主导影响度） 候选主导特征模式 $c=\{f_1,f_2,\cdots,f_k\}$ 的主导影响度 $DII(c)$ 定义为：

$$DII(c)=\frac{1}{|G|}\sum_{f_i\in G}\frac{1}{|\pi_{f_i}T(c)|}\sum_{i_i\in\pi_{f_i}T(c)}\frac{\sum_{f_j\in c,f_j\notin G}|\pi_{f_i}cl(i_i)|\dfrac{\min\{DV(i_i,f_k,c)\mid f_k\in c,f_k\notin G\}}{DV(i_i,f_j,c)}}{\sum_{f_j\in c,f_j\notin G}|\pi_{f_i}cl(i_i)|} \tag{4}$$

候选主导特征实例吸引其他特征实例的数量越多，且其他特征实例越靠近候选主导特征实例，则候选主导特征模式的主导影响度越大。

定义 6（主导特征模式和主导特征） 给定一个 k 阶候选主导特征模式 $c=\{f_1,f_2,\cdots,f_i,\cdots,f_k\}$ 及主导影响度阈值 min_dii，如果 $DII(c)\geqslant min\_dii$，则模式 c 称为主导特征模式，候选主导特征集 G 中的特征称为主导特征。

基于上述定义，本文所提主导特征模式挖掘问题定义如下。

问题定义：给定频繁 co-location 模式集合 FP，模式主导度阈值 min_pdi，主导影响度阈值 min_dii，主导特征模式挖掘就是从模式集合 FP 中发现所有满足下列条件的模式 c：

(1) $PDI(c)\geqslant min\_pdi$

(2) $DII(c)\geqslant min\_dii$

4 DFMASD 挖掘算法

本文所提的模式主导度和主导影响度不具有反单调性，不能利用这两个度量缩减模式搜索空间，因此主导特征模式挖掘的朴素方法是首先根据现有方法挖掘所有频繁式，然后根据定义 2 计算这些频繁模式主导度，筛选候选模式，根据定义 5 计算候选模式主导影响度，筛选主导特征模式。显而易见，朴素方法的挖掘效率较低。一方面，在参与度、主导度和主导影响度计算中，需要用到模式表实例，表实例的生成是较为耗时的步骤；另一方面，主导度和主导影响度本身的计算也比较耗时。为了提升挖掘效率，本文首先将频繁模式挖掘与主导特征模式挖掘相结合，即从 2 阶开始，逐阶依次判定频繁模式、候选模式、主导特征模式，以避免重复生成表实例；其次，提出引理 1 以优化主导度的计算，提出引理 2 优化主导影响度的计算。

引理 1 设模式 $c=\{f_1,f_2,\cdots,f_i,\cdots,f_k\}$ 中，参与率 $PR(f_i,c)$ 最小的两个特征为 f_{min1} 和 f_{min2} 且 $PR(f_{min1},c)\leqslant PR(f_{min2},c)$，参与实例数 $|\pi_{f_i}T(c)|$ 最大的两个特征为 f_{max1} 和 f_{max2} 且 $|\pi_{f_{max1}}T(c)|\geqslant|\pi_{f_{max2}}T(c)|$，则模式主导度的计算式式(2)可优化为：

$$PDI(c)=\begin{cases}FDP(f_{max1},c), & f_{min1}\neq f_{max1}\\ \min\{FDP(f_{max1},c),FDP(f_{max2},c)\}, & f_{min1}=f_{max1}\end{cases} \tag{5}$$

证明：1) 当 $f_{min1}\neq f_{max1}$ 时，因为对 $f_i\in c,f_i\neq f_{max1}$，有 $|\pi_{f_{max1}}T(c)|\geqslant|\pi_{f_i}T(c)|$，对 $f_i\in c,f_i\neq f_{min1}$，有 $PR(f_{min1},c)\leqslant PR(f_i,c)$，且 $f_{min1}\neq f_{max1}$，所以 $PDI(c)=\min\{FDP(f_i,c)\mid f_i\in c\}=\min\left\{\dfrac{\min\{PR(f_j,c)\mid f_j\in c,f_j\neq f_i\}}{|\pi_{f_i}T(c)|}\mid f_i\in c\right\}=\dfrac{PR(f_{min1},c)}{|\pi_{f_{max1}}T(c)|}=FDP(f_{max1},c)$。

2) 当 $f_{min1}=f_{max1}$ 时，因为对 $f_i\in c,f_i\neq f_{max1},f_i\neq f_{max2}$，有 $|\pi_{f_{max1}}T(c)|\geqslant|\pi_{f_{max2}}T(c)|\geqslant|\pi_{f_i}T(c)|$，对 $f_i\in c,f_i\neq f_{min1},f_i\neq f_{min2}$，有 $PR(f_{min1},c)\leqslant PR(f_{min2},c)\leqslant PR(f_i,c)$，且 $f_{min1}=f_{max1}$，所以 $PDI(c)=\min\{FDP(f_i,c)\mid f_i\in c\}=\min\left\{\dfrac{\min\{PR(f_j,c)\mid f_j\in c,f_j\neq f_i\}}{|\pi_{f_i}T(c)|}\mid f_i\in c\right\}=\min\left\{\dfrac{PR(f_{min2},c)}{|\pi_{f_{max1}}T(c)|},\dfrac{PR(f_{min1},c)}{|\pi_{f_{max2}}T(c)|}\right\}=\min\{FDP(f_{max1},c),FDP(f_{max2},c)\}$。

根据引理 1，我们仅需计算两个特征的主导率即可得到模式的主导度，从而减少计算量。

引理 2 在候选主导特征模式 c 中，对于 $\forall f_i\in G,\forall i_i\in\pi_{f_i}T(c)$，设 $f_{min}=\arg\min_{f_k\in c,f_k\notin G}\{DV(i_i,f_k,c)\}$，如果对于 $\forall f_j\in c,f_j\notin G,f_j\neq f_{min}$，$\dfrac{DV(i_i,f_{min},c)}{DV(i_i,f_j,c)}>min\_dii$，则 $DII(c)>min\_dii$，即候选主导特征模式 c 必为主导特征模式。

证明：式(4)中，因为

$$\frac{\sum_{f_j\in c,f_j\notin G}|\pi_{f_j}cl(i_i)|\dfrac{\min\{DV(i_i,f_k,c)\mid f_k\in c,f_k\notin G\}}{DV(i_i,f_j,c)}}{\sum_{f_j\in c,f_j\notin G}|\pi_{f_j}cl(i_i)|}=\frac{\sum_{f_j\in c,f_j\notin G}|\pi_{f_j}cl(i_i)|\dfrac{DV(i_i,f_{min},c)}{DV(i_i,f_j,c)}}{\sum_{f_j\in c,f_j\notin G}|\pi_{f_j}cl(i_i)|}$$

$$=\frac{\sum_{f_j\in c,f_j\notin G,f_j\neq f_{min}}|\pi_{f_j}cl(i_i)|\dfrac{DV(i_i,f_{min},c)}{DV(i_i,f_j,c)}+|\pi_{f_{min}}cl(i_i)|}{\sum_{f_j\in c,f_j\notin G}|\pi_{f_j}cl(i_i)|}$$

$$>\frac{\sum_{f_j\in c,f_j\notin G,f_j\neq f_{min}}|\pi_{f_j}cl(i_i)|*min\_dii+|\pi_{f_{min}}cl(i_i)|}{\sum_{f_j\in c,f_j\notin G}|\pi_{f_j}cl(i_i)|}$$

$$>min\_dii$$

所以 $DII(c)>min\_dii$。

利用引理 2，我们可以快速判断某些候选模式是否为主导特征模式，之后我们只需对不满足引理 2 的候选模式进一步根据定义 5 计算主导影响度，从而减少计算量。

本文所提主导特征模式挖掘算法 DFMASD（Pattern

with Dominant-Feature Mining Algorithm Based on Spatial Distribution)如算法 1 所示。

算法 1 DFMASD 算法

输入:空间实例集 S,空间特征集 F,距离阈值 d,参与度阈值 min_prev,主导度阈值 min_pdi,主导影响度阈值 min_dii

输出:主导特征模式集 SDCP

变量:k 为模式的阶;C_k 为 k 阶候选频繁模式集;P_k 为 k 阶频繁模式集,$T(c_k)$ 为 k 阶模式 c_k 的表实例

步骤:

1. $P_1=F$, $k=2$, $SDCP=\emptyset$;
2. WHILE($P_{k-1}\neq\emptyset$)
3. C_k=gen_candidate_patterns(P_{k-1});//生成候选频繁模式集
4. FOR EACH $c_k\in C_k$
5. $T(c_k)$=gen_table_instances(c_k,S,d);//生成模式表实例
6. IF Cal_PI(c_k,$T(c_k)$)≥min_prev //计算参与度并判断是否为频繁模式
7. $P_k\leftarrow c_k$;//加入频繁模式集
8. IF Cal_PDI(c_k,$T(c_k)$) ≥min_pdi //根据引理 1 计算主导度并判断是否为候选主导特征模式
9. IFCal_DII(c_k,$T(c_k)$)≥min_dii //根据引理 2 计算主导影响度并判断是否主导特征模式
10. SDCP←c_k;//加入主导特征模式集
11. END IF
12. END IF
13. END IF
14. END FOR
15. k=k+1;
16. END WHILE
17. RETURN SDCP.

步骤 1 进行初始化,步骤 2—16 从 2 阶开始,逐阶挖掘主导特征模式,其中步骤 3 基于 $k-1$ 阶频繁模式生成 k 阶候选频繁模式,对每个候选频繁模式,步骤 5 生成模式的表实例,步骤 6 计算模式参与度,并判断是否为频繁模式[6],若是,则步骤 8 根据引理 1 计算模式主导度,并判断是否为候选主导特征模式,若是,则步骤 9 根据引理 2 计算模式主导影响度,并判断是否为主导特征模式,若是,则步骤 10 将模式加入主导特征模式集中。

DFMASD 算法从 2 阶开始,逐阶依次生成频繁模式、候选主导特征模式、主导特征模式。对于给定空间数据集,频繁模式的数量受距离阈值和参与度阈值影响,候选主导特征模式的数量又受主导度阈值影响,主导特征模式的数量进一步受主导影响度阈值影响。生成 k 阶频繁模式中(步骤 3—7),主要时间开销是生成 $|C_k|$ 个 k 阶候选频繁模式 c_k 的表实例 $T(c_k)$,其时间复杂度为 $O(|C_k\|T(c_k)|)$[6]。基于 k 阶频繁模式的表实例及特征参与率,生成 k 阶候选主导特征模式(步骤 8)的主要时间开销是根据引理 1 计算 $|P_k|$ 个 k 阶频繁模式的主导度,其时间复杂度为 $O(k|P_k\|T(c_k)|)$。进一步生成 k 阶主导特征模式(步骤 9)的主要时间开销是引理 2 和定义 5 中对 $|CD_k|$ 个 k 阶候选主导特征模式的每一个模式,计算 $k-|G|$ 个其他特征的实例相对于 $|G|$ 个主导特征的实例的分布均值,其时间复杂度为 $O((k-|G|)|G\|CD_k\|T(c_k)|)$。由于 k, $|G|\leqslant|F|$,通常 $|F|\ll|T(c_k)|$, $|CD_k|<|P_k|<|C_k|$,因此 DFMASD 算法的时间复杂度为 $O(\sum_{k=2}^{|F|}|C_k\|T(c_k)|)$。

5 实验与分析

本节将通过实验分析所提算法 DFMASD 的运行效率和挖掘结果,并与文献[6]中的传统模式挖掘算法 join-based 和文献[20]中的主导特征模式挖掘算法 SK 进行比较。在实验中,算法采用 Java 语言实现,所有程序运行在 Intel Core I7 CPU、16 GB 内存、Windows 10 的 PC 上。

5.1 实验数据集及参数设置

为了分析在不同数据规模及分布的数据集上算法的运行效率,本文随机生成 3 个不同空间范围、不同特征数及实例数的合成数据集 Synthetic data 1, Synthetic data 2 和 Synthetic data 3。为了分析算法的挖掘结果,本文选用了 2 个真实数据集:"三江并流区域"植物数据集 Plantdata,其分布呈图 2 所示的带状分布;北京市 POI 数据集 Beijing-POI,其分布如图 3 所示。在植物数据集中,一类植物(如松茸)是一个特征;在 POI 数据集中,一类 POI(如停车场)是一个特征。这些数据集的描述信息如表 1 所列。

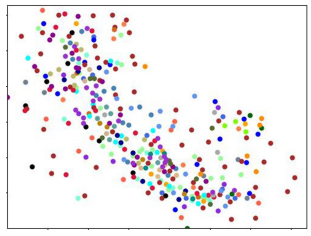


图 2 Plantdata 数据集分布

Fig. 2 Distribution of Plant data dataset

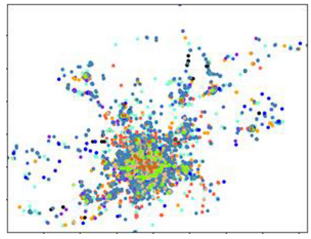


图 3 Beijing-POI 数据集分布

Fig. 3 Distribution of Beijing-POI dataset

表 1 数据集

Table 1 Dataset

数据集	特征数	实例数	空间范围
Synthetic data 1	10	10000	500×500
Synthetic data 2	10	10000	1000×1000
Synthetic data 3	25	20000	1000×1000
Plantdata	31	356	8000×13000
Beijing-POI	16	23025	22000×14000

在实验中,算法所需的参数距离阈值、参与度阈值、SK 算法的显著性阈值、DFMASD 算法的模式主导度阈值和主导影响度阈值在各个数据集上的默认设置如表 2 所列。

表 2 实验参数默认设置

Table 2 Default values of experimental parameters

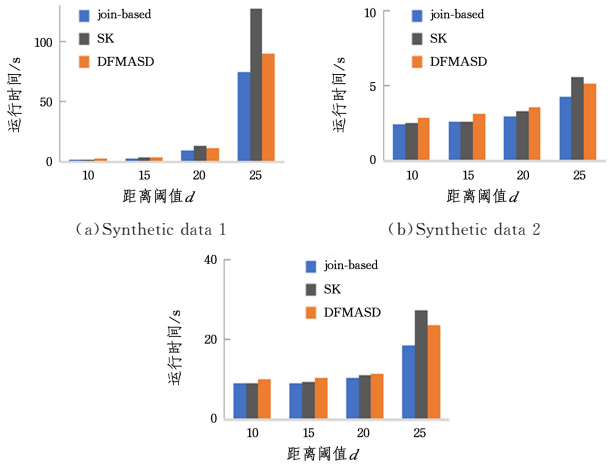
数据集	距离阈值	参与度阈值	显著性阈值	模式主导度阈值	主导影响度阈值
Synthetic data 1	25	0.20	0.4	0.0001	0.4
Synthetic data 2	25	0.20	0.1	0.0005	0.4
Synthetic data 3	25	0.40	0.1	0.0001	0.4
Plantdata	9000	0.25	0.2	0.1500	0.5
Beijing-POI	50	0.25	0.2	0.0003	0.5

5.2 算法运行效率的分析

我们首先分析距离阈值和参与度阈值对 3 种算法 join-based，SK 和 DFMASD 运行效率的影响，然后分析模式主导度阈值和主导影响度阈值对本文所提 DFMASD 算法运行效率的影响。

5.2.1 距离阈值对算法运行时间的影响

距离阈值 d 分别取 10，15，20，25 时，3 种算法在合成数据集上的运行时间如图 4 所示。



(a) Synthetic data 1　(b) Synthetic data 2

(c) Synthetic data 3

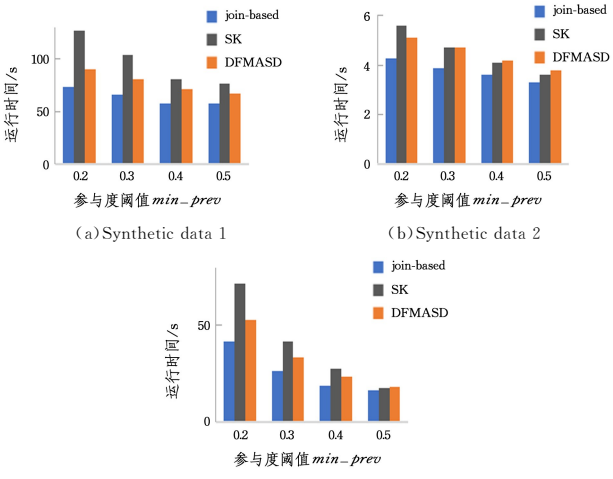
图 4　距离阈值对算法运行时间的影响

Fig. 4　Influence of d on running time of algorithm

从图 4 可以看出，在 3 个合成数据集上，3 种算法的运行时间均随着距离阈值 d 的增大而增加，这是由于距离阈值 d 的增大导致形成团的行实例和频繁模式增多，从而算法的运行时间增加。DFMASD 算法的运行时间大致为 join-based 算法运行时间的 1.2 倍，这是因为主导特征模式挖掘是在频繁模式挖掘的基础上进行的。当距离阈值为 25 时，DFMASD 算法的运行时间比 SK 算法的运行时间短，这是由于 DFMASD 算法采用了主导度筛选和影响度筛选两次筛选，并且采用了引理 1 和引理 2 优化主导度和影响度的计算，当频繁模式增加时，优化效果更为明显。

5.2.2 参与度阈值对算法运行时间的影响

参与度阈值分别取 0.2，0.3，0.4，0.5 时，3 种算法在合成数据集上的运行时间如图 5 所示。



(a) Synthetic data 1　(b) Synthetic data 2

(c) Synthetic data 3

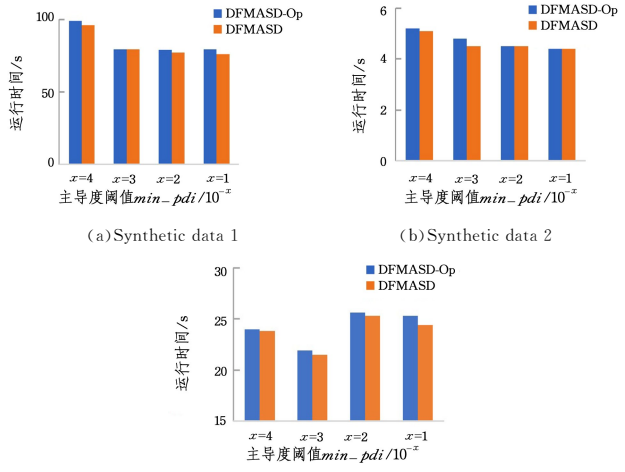
图 5　参与度阈值对算法运行时间的影响

Fig. 5　Influence of *min_prev* on running time of algorithm

从图 5 可以看到，在 3 个合成数据集上，3 种算法的运行时间均随着参与度阈值增大而减少，这是由于参与度阈值增大导致频繁模式减少，从而算法的运行时间缩短。当参与度阈值增大到一定程度时，由于频繁模式数量大幅减少，3 种算法的运行时间差距缩小。

5.2.3 主导度阈值对 DFMASD 算法运行时间的影响

在 3 个合成数据集上，我们分析主导度阈值分别取 10^{-4}，10^{-3}，10^{-2}，10^{-1} 时，DFMASD 算法的运行效率，并与去除 DFMASD 算法中两个度量的优化计算得到的算法 DFMASD-Op 进行比较，分析引理 1 和引理 2 对算法运行时间的影响，实验结果如图 6 所示。



(a) Synthetic data 1　(b) Synthetic data 2

(c) Synthetic data 3

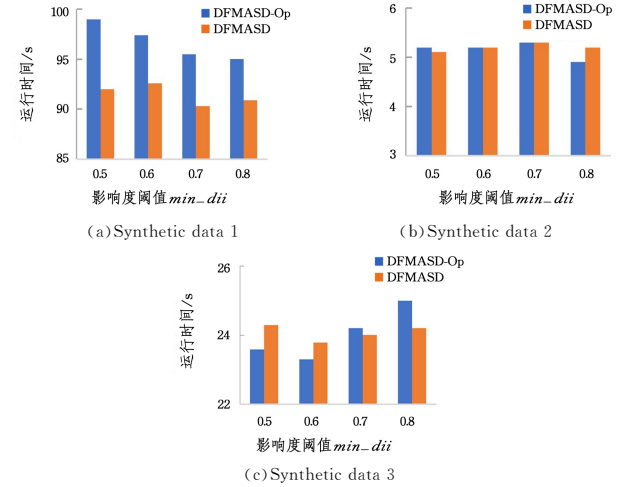
图 6　主导度阈值对 DFMASD 算法运行时间的影响

Fig. 6　Influence of *min_pdi* on running time of algorithm

图 6 给出了在 3 个合成数据集上，DFMASD-Op 算法和 DFMASD 算法的运行时间。因为利用引理进行了算法优化，所以在每个数据集上，DFMASD 的总体运行时间要短于 DFMASD-Op 的运行时间。

5.2.4 影响度阈值对 DFMASD 算法的影响

类似地，在 3 个合成数据集上，我们分析影响度阈值分别取 0.5，0.6，0.7，0.8 时，DFMASD 算法的运行效率，并与算法 DFMASD-Op 进行比较，实验结果如图 7 所示。



(a) Synthetic data 1　(b) Synthetic data 2

(c) Synthetic data 3

图 7　影响度阈值对 DFMASD 算法运行时间的影响

Fig. 7　Influence of *min_dii* on running time of algorithm

从图 7 可以看到，在某些数据集上，DFMASD 算法的

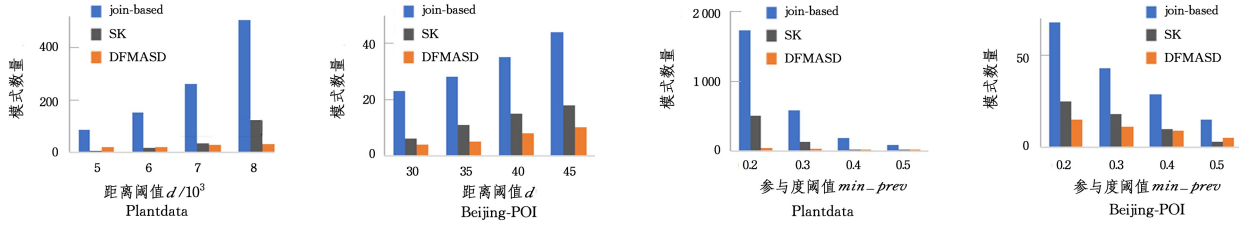
运行时间比 DFMASD-Op 算法的运行时间长，这是因为当不能利用引理 2 进行算法优化时，需要根据影响度阈值的定义进行主导特征模式判定，增加了算法运行时间。在较为稠密的 Synthetic data 1 上，利用引理的算法运行时间总是比不利用引理的算法运行时间短，可见引理在较为稠密的数据集上效果较好，这是由于在稠密数据集上，频繁模式数量增多，会有更多模式可以利用引理 1 和引理 2 进行算法优化。

5.3 算法挖掘结果分析

在两个真实数据集上，分析比较所提 DFMASD 算法与文献[20]中 SK 算法以及文献[6]中 join-based 算法的挖掘结果。

5.3.1 距离阈值和参与度阈值对模式数量的影响

图 8 给出了在不同距离阈值和参与度阈值下 DFMASD 算法、SK 算法和 join-based 算法在真实数据集上挖到的模式数量。



(a)距离阈值对模式数量的影响　　(b)参与度阈值对模式数量的影响

图 8　距离阈值和参与度阈值对模式数量的影响

Fig. 8　Influence of d and min_prev on number of patterns

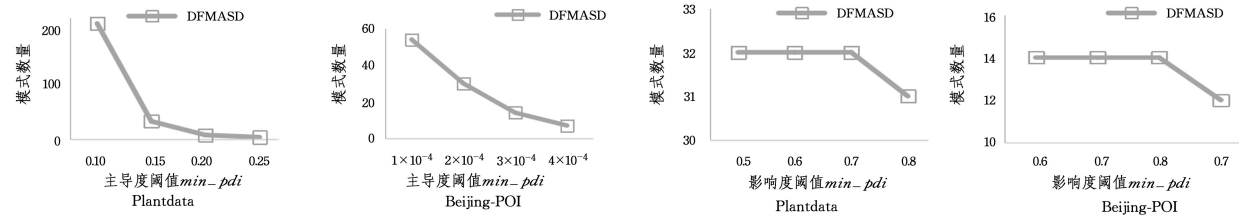
从图 8 可以看出，距离阈值和参与度阈值的变化影响了频繁模式的数量，进而影响了主导特征模式的数量。

在 Plantdata 数据集上，与 SK 算法相比，DFMASD 算法的模式数量受距离阈值变化的影响较小，这是因为 DFMASD 算法的主导影响度考虑了实例的空间分布，在较稀疏的 Plantdata 数据集上，即使距离阈值增大，主导影响度也变化不大，从而导致模式数量变化不大。

在 Beijing-POI 数据集上，两种算法的模式数量都受到距离阈值和参与度阈值的影响。当距离阈值和参与度阈值变化不大时，SK 算法的模式数量较 DFMASD 算法变化明显，这是因为 DFMASD 算法综合考虑特征的参与实例数和实例间的距离，当团中实例数增多或减少时，DFMASD 算法都能综合评估模式中特征的主导强度。

5.3.2 主导度阈值和影响度阈值对模式数量的影响

图 9 给出了在不同主导度阈值和影响度阈值下 DFMASD 算法挖到的模式数量。从图 9 可以看到，DFMASD 算法挖到的模式数量随主导度阈值和影响度阈值的增加而减少。



(a)主导度阈值对模式数量的影响　　(b)影响度阈值对模式数量的影响

图 9　主导度阈值和影响度阈值对模式数量的影响

Fig. 9　Influence of min_pdi and min_dii on number of patterns

在 Plantdata 数据集上，DFMASD 算法挖到的模式数量受主导度影响较大，受影响度阈值影响较小，这是由于 Plantdata 数据集是较为稀疏的数据集，模式中特征实例参与较少，容易受到主导度阈值的影响。影响度阈值考察参与实例分布状态，整体受到影响度阈值的影响较小。

在 Beijing-POI 数据集上，DFMASD 算法挖掘的模式数量受主导度和影响度影响较小。这是由于 Beijing-POI 数据集是较为稠密的数据集，模式中特征参与率较高，当主导度阈值升高时，候选主导特征模式减少较少，从而主导特征模式减少较缓慢。由于模式中特征参与实例数较多，特征所表现出的主导性也较强，受影响度阈值影响较小。

5.4 主导特征模式实例分析

在两个真实数据集上，我们对本文所提 DFMASD 算法与文献[20]中 SK 算法挖掘到的模式进行实例分析。

首先，由于 SK 算法通过特征在模式及其子模式中的参与率变化来度量特征及模式的显著性，因此 SK 算法只能挖掘到 3 阶及以上的主导特征模式。然而，DFMASD 算法可以挖掘到 2 阶及以上的主导特征模式，而且这些 2 阶主导特征模式除了反映两个特征间的主导关系外，在一定程度上也可以帮助我们理解高阶主导特征模式。例如，在 Plantdata 数据集上，“丽江雪胆”是 2 阶模式{丽江雪胆 *，金荞麦}、{丽江雪胆 *，高河菜}的主导特征，那么不难理解在 3 阶模式{丽江雪胆 *，高河菜，延龄草}、{丽江雪胆 *，金荞麦，延龄草}中它也是主导特征。同理，在 Beijing-POI 数据集上，2 阶主导特征模式{招待所 *，咖啡屋}也可以帮助我们理解 3 阶主导特征模式{招待所 *，咖啡屋，停车场}和{招待所 *，咖啡屋，服装店}。

其次，DFMASD 算法与 SK 算法挖到的主导特征模式既有相同的模式，也有不同的模式，表 4 列出了部分代表性模式。从表 4 中可以看到，两种算法都挖到了 3 阶模式{招待所 *，中餐，服装店}；而 DFMASD 算法挖到了 SK 算法没有挖到的 3 阶模式{招待所 *，咖啡屋，服装店}和{宾馆 *，咖啡屋，招待所}，这两种模式较好地反映了现实情况：招待所附近通常会有咖啡屋和服装店这类生活便利小店，而宾馆附近除了咖啡屋外，通常还会有更为经济便利的招待所；同时 SK 算法也挖到了 DFMASD 算法没有挖到的 3 阶模式{咖啡屋 *，招待所，宾馆}和{服装店 *，招待所，宾馆}，然而这两种模式与现实情况存在差异，通常人们不会因为咖啡屋和服装店而在附近建招待所和宾馆。4 阶模式也有类似的情况。从表 4 中还

可以看到，两个算法都判定 5 阶模式{宾馆，中餐，招待所，停车场，服装店}是主导特征模式，但是它们判定的主导特征不同，DFMASD 算法判定宾馆主导其他特征，SK 算法判定招待所主导其他特征，根据现实情况，宾馆作为主导特征更为合理。我们还发现，在 3 阶模式中也有类似的情况。DFMASD 算法挖到的模式{宾馆 *，咖啡屋，招待所}和 SK 算法挖到的模式{咖啡屋 *，招待所，宾馆}相比，模式中的特征都相同，但是识别的主导特征不同。从现实情况来看，DFMASD 算法识别到的主导特征更为合理。除此之外，SK 算法识别的主导特征变动较大，即其识别的主导特征易受其他特征影响。

表 4　DFMASD 算法与 SK 算法挖到的部分主导特征模式

Table 4　Some dominant feature patterns mined by DFMASD and SK algorithms

模式	DFMASD 算法	SK 算法
2 阶	**{花园 *，咖啡屋}** **{招待所 *，咖啡屋}**	
3 阶	{招待所 *，中餐，服装店} **{招待所 *，咖啡屋，服装店}** **{宾馆 *，咖啡屋，招待所}**	{招待所 *，中餐，服装店} *{咖啡屋 *，招待所，宾馆}* *{服装店 *，招待所，宾馆}*
4 阶	{招待所 *，中餐，停车场，服装店} {宾馆 *，中餐，招待所，服装店}	{招待所 *，中餐，停车场，服装店} {服装店 *，招待所，宾馆，停车场}
5 阶	**{宾馆 *，中餐，招待所，停车场，服装店}**	*{招待所 *，宾馆，中餐，停车场，服装店}*

注：粗体为 DFMASD 算法挖到而 SK 算法没有挖到的模式；斜体为 SK 算法挖到而 DFMASD 算法没有挖到的模式

结束语　针对空间 co-location 模式中的主导关系，本文研究了主导特征的度量方法及主导特征模式的挖掘方法。分析模式中特征实例的空间分布，综合特征间主导关系的可能倾向和影响强度，提出模式主导度和主导影响度用以识别主导特征及模式，进而通过优化新度量的计算提出有效的主导特征及模式挖掘算法。在合成数据集和真实数据集上的大量实验验证了所提方法能够挖掘合理、实用、有价值的主导特征模式。在今后的工作中，我们将设计有效的数据结构和算法以减少表实例的存储空间，并提高算法的运行效率。

参 考 文 献

[1] TOBLER W R. A Computer Movie Simulating Urban Growth in the Detroit Region[J]. Economic Geography, 2016, 46: 234-240.

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

[3] AKBARI M, SAMADZADEGAN 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.

[4] AKBARI M, SAMADZADEGAN 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.

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

[6] HUANG Y, SHEKHAR S, XIONG H. Discovering colocation patterns from spatial data sets: a general approach[J]. IEEE Transactions on Knowledge & Data Engineering, 2004, 16(12): 1472-1485.

[7] JIN S Y, SHEKHAR S. A partial join approach for mining co-location patterns[C]// 12th ACM International Workshop on Geographic Information Systems. Washington, DC, USA, ACM, 2004.

[8] JIN S Y, SHEKHAR S, CELIK M. A join-less approach for co-location pattern mining: a summary of results[C]// IEEE International Conference on Data Mining. IEEE, 2005.

[9] WANG L, BAO Y, LU J, et al. A new join-less approach for co-location pattern mining[C]// IEEE International Conference on Computer & Information Technology. IEEE, 2008.

[10] WANG L, BAO Y, LU Z. Efficient Discovery of Spatial Co-Location Patterns Using the iCPI-tree[J]. Open Information Systems Journal, 2009, 3(2): 69-80.

[11] WANG L, ZHOU L, LU J, et al. An order-clique-based approach for mining maximal co-locations[J]. Information Sciences, 2009, 179(19): 3370-3382.

[12] 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.

[13] YUAN F, WANG L, TENG H. Spatial Co-location Pattern Mining Based on Density Peaks Clustering and Fuzzy Theory[C]// Asia-Pacific Web(APWeb) and Web-Age Information Management(WAIM) Joint International Conference on Web and Big Data. Cham: Springer, 2018.

[14] LEI L, WANG L Z, XIAO Q. Study on fuzzy mining technology in spatial co-location pattern mining[J]. Computer Engineering and Applications, 2019, 55(21): 158-166.

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

[16] WANG X, WANG L, LU J, et al. Effectively Updating High Utility Co-location Patterns in Evolving Spatial Databases[M]. Springer International Publishing, 2016.

[17] HUANG Y, PEI J, XIONG H. Mining Co-Location Patterns with Rare Events from Spatial Data Sets[J]. GeoInformatica, 2006, 10(3): 239-260.

[18] FENG L, WANG L Z, GAO S J. A new approach of mining co-location patterns in spatial datasets with rare features[J]. Journal of Nanjing University(Natural Sciences), 2012, 48(1): 99-107.

[19] YUAN F, WANG L, WANG X, et al. Mining Co-location Patterns with Dominant Features[C]// International Conference on Web Information Systems Engineering. 2017.

[20] FANG Y, WANG L Z, ZHOU L H. Mining Spatial Co-location Patterns with Key Features[J]. Journal of Data Acquisition and Processing, 2018, 33(4): 692-703.

[21] MA D, CHEN H M, WANG L Z, et al. Dominant feature mining of spatial sub-prevalent co-location patterns[J]. Journal of Computer Applications, 2020, 40(2): 465-472.

XIONG Kai-fang, born in 1993, master. His main research interests include spatial data mining and so on.

CHEN Hong-mei, born in 1976, Ph. D, associate professor, is a member of China Computer Federation. Her main research interests include spatial data mining and so on.