Computer Science and Application 计算机科学与应用, 2021, 11(1), 176-194
Published Online January 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.111019

Hans汉斯

基于模糊邻近关系挖掘含主导特征的空间并置模式

冯　时，王丽珍，方　圆

云南大学信息学院，云南　昆明
Email: 954787012@qq.com, lzhwang@ynu.edu.cn

收稿日期：2020年12月27日；录用日期：2021年1月21日；发布日期：2021年1月28日

摘　要

空间并置(co-location)模式挖掘旨在发现空间中频繁在一起出现的空间特征的子集。空间并置模式中有一类模式其特征的地位是不平等，发现含主导特征的并置模式可以为实际应用提供更为精准的决策支持。由于单一的邻近距离阈值判定两个空间实例间的邻近性会导致邻近关系的缺失，因此，本文首先定义空间实例间的模糊邻近关系，然后定义模式中特征的模糊影响度和模糊影响比识别含主导特征的并置模式；其次，提出基于模糊邻近关系的含主导特征的并置模式挖掘算法及算法优化策略；最后，在合成数据集和真实数据集上验证了算法的正确性和有效性，并在真实数据集上对挖掘结果的实用性进行了比较和分析。

关键词

空间数据挖掘，空间并置模式，模糊邻近关系，主导特征，主导特征模式

Mining Spatial Collocation Patterns with Dominant Features Based on Fuzzy Neighborhood Relationship

Shi Feng, Lizhen Wang, Yuan Fang

School of Information, Yunnan University, Kunming Yunnan
Email: 954787012@qq.com, lzhwang@ynu.edu.cn

Received: Dec. 27th, 2020; accepted: Jan. 21st, 2021; published: Jan. 28th, 2021

文章引用: 冯时, 王丽珍, 方圆. 基于模糊邻近关系挖掘含主导特征的空间并置模式[J]. 计算机科学与应用, 2021, 11(1): 176-194. DOI: 10.12677/csa.2021.111019

Abstract

Spatial co-location pattern mining aims at mining the collection of spatial features that are frequently occurring together in a space. There is a kind of pattern in spatial co-location patterns whose feature position is inequality. Finding co-location patterns with dominant features can provide more accurate decision support for practical applications. A single distance threshold determines the neighborhood relationship between two spatial instances could lead to a lack of the neighborhood relationship. So, firstly, a fuzzy neighbor relation between spatial instances is defined, and then the fuzzy influence degree and influence ratio of the features in a co-location pattern are defined to identify the co-location pattern with dominant features. Secondly, based on fuzzy neighborhood relationship, a mining algorithm and an optimization strategy of co-location patterns with dominant features are proposed. Finally, the correctness and effectiveness of the algorithm are verified on the synthetic and real data sets, and the practicability of mining results on the real data sets is compared and analyzed.

Keywords

Spatial Data Mining, Spatial Co-Location Pattern, Fuzzy Neighborhood Relationship, Dominant Feature, Dominant Feature Pattern

1. 引言

随着基于位置服务(LBS)、全球定位系统和移动电子设备的快速发展，带有空间位置信息的数据急速增长，产生了大量的空间数据[1]。空间数据挖掘旨在从海量、高维的空间数据中挖掘潜在有用的和有价值的信息[2]。空间并置(co-location)模式挖掘作为空间数据挖掘的一个重要研究方向，在环境保护[3]、城市计算[4]、公共交通[5]等领域具有重要和广泛的应用。空间并置模式是一组空间特征的子集，它们的实例在邻域内频繁并置出现。例如，医院附近往往存在药店和花店，根瘤菌往往长在豆科植物旁等等。

传统的空间并置模式挖掘一般采用最小参与率(参与度)度量，即用一组空间特征的实例在邻域中共同出现的频率衡量其模式的有趣程度[6]。然而，基于参与度的频繁并置模式挖掘框架无法得到模式中特征的主导关系。识别空间并置模式中特征之间的主导关系，可以更好地为空间并置模式挖掘的实际应用提供决策支持。例如在植被数据上挖掘出的{松树，杉树，松茸}是一个频繁空间并置模式，其参与度可以很好地反映出该特征[7]组共现的强度，但参与度信息不能体现“松茸”的生存受到“松树”和“杉树”的影响，即在这个模式中“松树”和“杉树”是这个模式的主导特征。在植物分布分析和应用研究中，尽管挖掘出的频繁并置模式可以发现共生的植物物种关系，但是为了进一步研究植物群落和分布的结构和特征，挖掘出含主导特征的空间并置模式可以为植物学家提供更丰富的信息。另外，含主导特征的并置模式挖掘也可以为商业选址和主导设施的建立提供重要的方向和信息。然而，目前提出的含主导特征的空间并置模式挖掘方法没有考虑到空间特征实例的邻近关系是一个模糊的概念，用单一的邻近距离阈值确定两个空间实例的邻近关系会造成邻近关系的缺失；另外，在定义模式中特征贡献度和影响度时也没有考虑到贡献和影响程度的模糊性。

基于以上思考，本文将实例的邻近关系作为模糊概念定义了模糊邻近关系，基于此定义了实例的模

糊行实例贡献度、实例的模式贡献度、模糊影响度和模糊影响比等重要概念，提出基于模糊邻近关系的含主导特征的空间并置模式挖掘问题。并针对以下两方面的挑战：1) 如何将模糊邻近关系引入到空间并置模式中并度量特征间主导关系；2) 如何基于模糊邻近关系提出高效地挖掘含有主导特征的空间并置模式算法，进行研究。主要贡献包括：

1) 基于模糊集理论给出模糊邻近关系的定义，分析所给出的模糊邻近关系具有的性质。

2) 定义实例的模糊贡献度以度量实例在模式中的贡献程度，并基于此定义模糊影响度和模糊影响比以度量特征在模式中的影响(主导)程度。

3) 设计了含主导特征的空间并置模式挖掘算法 DFMAFPR 和优化算法 DFMAFPR-Improved，并在合成数据集和真实数据集上进行大量的实验，验证算法的有效性。

2. 相关工作

空间并置模式挖掘在 Huang [8]等人的文献中针对确定数据提出了一个基于完全连接的经典算法(称为 join-based 算法)。为提高挖掘效率，各种改进算法如基于星型邻居物化的 Join-less 算法[9]、基于前缀树物化的 CPI-tree 算法[10]和 iCPI-tree 算法[11]、基于有序团物化的 order-clique-based 算法[12]等相继被提出。针对不确定数据，Wang 等提出了概率频繁空间并置模式挖掘算法[13] [14]，区间数据上的空间并置模式挖掘算法[15]。针对模糊数据，文献[16]提出了模糊参与率和模糊参与度概念，设计了有效的模糊空间并置模式挖掘算法；文献[17]通过将模糊理论和密度峰值聚类算法相结合，通过聚类对空间数据进行分类，并采用模糊团的概念挖掘并置模式；文献[18]基于模糊理论定义了特征邻近度并采用模糊聚类算法挖掘并置模式。针对带约束的空间数据，文献[19]提出了最大参与率概念并设计了 maxPrune 算法挖掘带有稀有特征的并置模式；文献[20]针对 maxPrune 算法的问题提出了最小加权参与率更有效地挖掘带稀有特征的并置模式。针对含主导特征的空间并置模式挖掘问题，文献[21]通过计算模式特征在频繁模式及其子模式的特征参与率变化来判断特征在模式中的影响程度[22]，并根据模式中特征之间的影响度差异定义了特征差异度作为特征在模式中的主导程度的度量指标来识别含主导特征的 co-location 模式及其主导特征。然而，现有的主导特征模式挖掘算法没有考虑到空间邻近关系的模糊性[23]，主导特征模式在挖掘实际交通中交通堵塞问题[24]时会造成邻近关系的缺失[25]，在计算频繁并置模式时会丢失一些有价值的模式，继而在挖掘主导特征时挖掘不到一些有参考价值的特征。

本文在传统并置模式主导特征挖掘方法基础上，考虑空间特征实例间的模糊邻近关系，研究了基于模糊邻近关系的含主导特征并置模式挖掘问题。

3. 相关定义

空间并置模式[1]：给定一个空间特征集 $O=\{o_1,o_2,\cdots,o_n\}$，对应空间实例集合 $S=S_1\cup S_2\cup\cdots\cup S_n$，其中 S_i 是特征 o_i $(1\le i\le n)$的实例集合，以及距离阈值 d。如果两个实例 $s_i, s_j\in S$ 的欧几里得距离小于等于距离阈值 d，则称两个实例满足邻近关系 R，即 $R(s_i, s_j)$。一个空间特征集 O 的 k 阶子集 $c=\{o_1,o_2,\cdots,o_k\}$ $(c\subseteq O,\ k=|c|)$称为 k 阶并置模式。为衡量 c 的频繁程度，引入模式的行实例和表实例：若实例集 $I=\{s_1,s_2,\cdots,s_k\}$ $(I\subseteq S)$包含 c 中所有特征且 I 中的任一子集都不包含 c 的所有特征，并且 I 在邻近关系 R 下形成团关系，则称 I 为 c 的一个行实例，记为 $R(c)$，c 的所有行实例组成 c 的表实例，记为 $T(c)$。

在空间并置模式中特征的参与率 $PR(c, o_u)$定义为模式 c 中特征 o_u 的实例在 c 的表实例中不重复出现的个数与 o_u 总实例个数之比，即

$$PR(c,o_u)=\frac{\left|\pi_{o_u}(T(c))\right|}{\left|T(o_u)\right|} \tag{1}$$

其中 π 是关系的投影操作。

模式 c 的参与度 $PI(c)$定义为模式 c 中所有特征参与率的最小值，即

$$PI(c)=\min_{u=1}^{k}\{PR(c,o_u)\} \tag{2}$$

给定参与度阈值 min_$prev$，当 $PI(c)\geq \min\_prev$，则称模式 c 为频繁并置模式。

如图 1 所示，图中有三个空间特征 A、B 和 C，设距离阈值 $d=100$，参与度阈值 $\min\_prev=0.3$，则实例 A_1 和 B_1 满足邻近关系 $R(A_1, B_1)$，模式{A, B, C}的表实例 $T(\{A,B,C\})=\{\{A_1,B_1,C_2\},\{A_3,B_2,C_3\},\{A_4,B_4,C_2\}\}$，模式{A, B, C}的参与度为 $PI(\{A,B,C\})=\min(0.43,0.5,0.5)=0.42$，则模式{A, B, C}是一个频繁并置模式。

定义 1 模糊邻近关系(FNR)：以空间实例集 S 中两两实例间的欧氏距离 D 作为论域 $D=[0,\infty)$，模糊邻近关系(FNR)是基于距离 D 的邻近关系的集合。空间任意两个实例 s_i 和 s_j 间的欧氏距离记为 d，给定下面映射关系：FNR: $D\rightarrow[0, 1]$, $d\rightarrow\mu(s_i, s_j)$，则称 μ 确定了 D 上的一个模糊子集 FNR，μ 为 FNR 的隶属函数，$\mu(s_i, s_j)$为距离 d 对 FNR 的隶属度，也称邻近度。空间实例集 S 中的任意两个实例 s_i 和 s_j 的模糊邻近关系 FNR 表示为：

$$\mathrm{FNR}=\left\{\mu(s_i,s_j)\middle|s_i,s_j\in S\right\} \tag{3}$$

给定用户自定义距离阈值 d_1, d_2，其中 $\mu(s_i, s_j)$定义为：

$$\mu(s_i,s_j)=\begin{cases}1, & d\leq d_1\\ 1-\dfrac{d-d_1}{d_2-d_1}, & d_1<d\leq d_2\\ 0, & d>d_2\end{cases} \tag{4}$$

例 1：空间特征 A、B 和 C 的实例分布如图 1 所示，给定距离阈值 $d_1=100$, $d_2=300$，选取任意两个空间实例 A_1 和 B_1，假设 A_1 和 B_1 的欧式距离 $\mathrm{dist}(A_1, B_1)=140$，则 A_1 和 B_1 的模糊邻近度 $\mu(A_1, B_1)=0.8$。

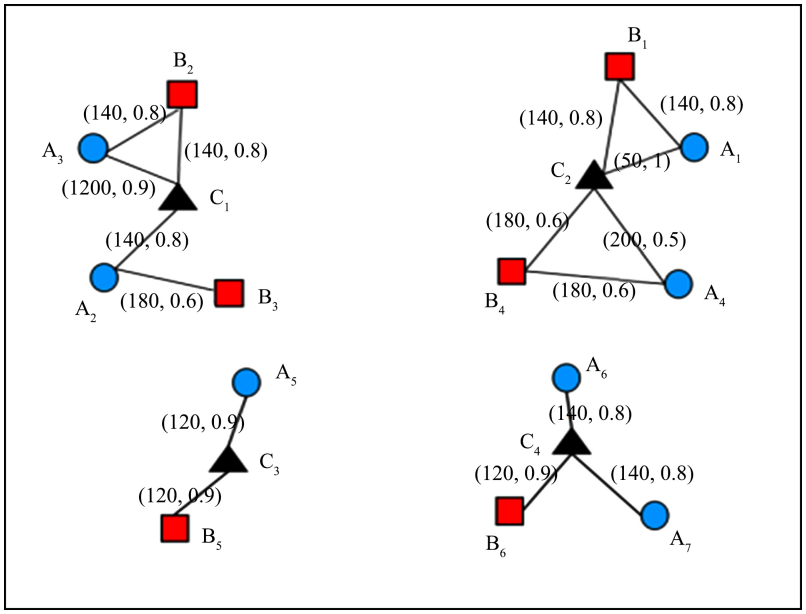

Figure 1. Spatial features and their instance distribution
图 1. 空间特征及其实例分布

定义 2 FNR 的 α-截集：给定用户自定义的邻近度阈值 α，模糊邻近关系 FNR 的 α-截集 FNR_α 定义为实例 s_i 和 s_j 邻近度 $\mu(s_i, s_j)$不小于 α 的 FNR 的子集，即

$$\mathrm{FNR}_\alpha = \left\{ \mu\left(s_i, s_j\right) \middle| s_i, s_j \in S, \mu\left(s_i, s_j\right) \geq \alpha \right\} \tag{5}$$

其中 $\mu(s_i, s_j) \geq \alpha$ 表示为实例 s_i 和 s_j 满足 α 邻近关系，也可表示为 $\mu_\alpha(s_i, s_j)$。

例 2：给定邻近度阈值 $\alpha = 0.1$, A_1 和 B_1 的邻近度为 0.8，那么空间实例 A_1 和 B_1 满足 α 邻近关系，即 $\mu_{0.1}(\mathrm{A}_1, \mathrm{B}_1)$。

基于模糊邻近关系的并置模式 c 的一个模糊行实例 FR 是空间实例集，即 $FR \subseteq S$。FR 具备以下特征：1) 在 α 邻近关系下形成团；2) FR 包含了模式 c 中的所有特征；3) 没有任意一个 FR 的子集可以包含 c 中所有的特征。模式 c 的**模糊表实例是**所有模糊行实例的集合，记为 $FT(c) = \{FR_1, FR_2, \cdots, FR_m\}$。

例 3：如图 1 空间实例集中，$\{\mathrm{A}_1, \mathrm{B}_1, \mathrm{C}_2\}$是并置模式{A, B, C}的一个模糊行实例，模式{A, B, C}的模糊表实例为 $\mathrm{FT}(\{\mathrm{A,B,C}\}) = \{\{\mathrm{A}_1,\mathrm{B}_1,\mathrm{C}_2\},\{\mathrm{A}_3,\mathrm{B}_2,\mathrm{C}_3\},\{\mathrm{A}_4,\mathrm{B}_4,\mathrm{C}_2\}\}$。

定义 3 模糊星型邻居(FSN)：给定一个度量实例模糊邻近关系的隶属函数 μ，一个空间实例 s，它的特征类型是 $o \in O$，实例 s 的模糊星型邻居定义为它本身和其所有特征类型大于它的模糊邻居的邻近度大于等于 α 的空间实例的集合，即

$$\mathrm{FSN}(s) = \left\{ \left(s_i \middle| = s \vee \mu\left(s, s_i\right) \geq \alpha \right) \right\} \tag{6}$$

其中，s 称为中心实例。

根据图 1 中空间实例分布及实例间模糊邻近度，可以列出如图 2 的空间特征 A 和 B 的实例的模糊星型邻居集。

中心		模糊邻居实例	中心		模糊邻居实例
特征	实例		特征	实例	
A	A_1	$\mathrm{A}_1,\mathrm{B}_1,\mathrm{C}_2$	B	B_1	$\mathrm{B}_1,\mathrm{C}_2$
	A_2	$\mathrm{A}_1,\mathrm{B}_1,\mathrm{C}_2$		B_2	$\mathrm{B}_2,\mathrm{C}_2$
	A_3	$\mathrm{A}_1,\mathrm{B}_1,\mathrm{C}_2$		B_3	B_3
	A_4	$\mathrm{A}_1,\mathrm{B}_1,\mathrm{C}_2$		B_4	$\mathrm{B}_4,\mathrm{C}_2$
	A_5	C_3		B_5	$\mathrm{B}_5,\mathrm{C}_3$
	A_6	C_4		B_6	$\mathrm{B}_6,\mathrm{C}_4$
	A_7	C_4			

Figure 2. The star neighbor sets of features A and B in Figure 1
图 2. 图 1 中特征 A, B 的星型邻居集

传统并置模式挖掘中，使用模式中特征参与率的最小值作为模式的参与度，进而来衡量模式的频繁性。在频繁模式的一个行实例中两两实例都满足距离阈值 d，某一个特征的一个实例在模式中的参与率的贡献为“1”。没有出现在行实例中的实例贡献为“0”。但是实际中，实例对于特征参与率的贡献是一个模糊的概念，为此，引入下面实例的模糊行实例贡献度和实例的模式贡献度概念。

定义 4 实例的模糊行实例贡献度：给定一个 k 阶并置模式 c，及其模糊表实例 $FT(c)=\{FR_1,FR_2,\cdots,FR_m\}$，其任一模糊行实例 $FR=\{s_1,s_2,\cdots,s_k\}$ （$FR\in FT(c)$），对于 FR 中的任一实例 $s_i\in FR$，s_i 在模糊行实例 FR 中的贡献定义为 s_i 与 FR 中其它实例之间的模糊邻近度的最小值，即

$$FR\_\mathrm{Contri}(FR,s_i)=\underset{s_j\in FR,i\neq j}{\mathrm{Min}}\left(\mu(s_i,s_j)\right) \tag{7}$$

由于空间的连续性，相同的实例往往会参与在不同的行实例中，因此出现在多个模糊行实例中的同一实例将出现多个贡献度，为此，我们给出**实例的模式贡献度。**

定义 5 实例的模式贡献度：给定一个 k 阶并置模式 c 及其模糊表实例 $FT(c)=\{FR_1,FR_2,\cdots,FR_m\}$，对任一实例 $s_i\in FT(c)$，s_i 在 $FT(c)$ 中所属行实例集定义为 $FT\_s_i(c)=\{FR_1,FR_2,\cdots,FR_h\}$（$m\le h$，$FT\_s_i(c)\subseteq FT(c)$），则 s_i 对模式的贡献度定义为 s_i 对其所属模糊行实例的贡献的最大值，即：

$$Contri(FT(c),s_i)=\underset{FR\in FT\_s_i(c)}{\mathrm{Max}}\left(FR\_Contri(FR,s_i)\right) \tag{8}$$

例 4：图 1 中并置模式{A, B, C}的模糊行实例{A_1, B_1, C_2}中实例 C_2 的模糊行实例贡献度为 $FR\_Contri(\{A_1,B_1,C_2\},C_2)=\min\{m(A_1,C_2),m(B_1,C_2)\}=\min\{0.8,1\}=0.8$。实例 C_2 的模式贡献度为 $Contri(\{\{A_1,B_1,C_2\},\{A_4,B_4,C_2\}\},C_2)=\mathrm{Max}\{0.8,0.5\}=0.8$。

相对于传统并置模式的参与率和参与度，基于实例在模糊行实例中的贡献定义模糊参与率和模糊参与度以衡量模式的频繁性。

定义 6 模糊参与率(FPR)和模糊参与度(FPI)：给定一个并置模式 $c=\{o_1,o_2,\cdots,o_k\}$，其空间特征 $o_i\in c$ $(1\le i\le k)$的模糊参与率定义为 o_i 参与在 c 中的所有实例对模式的贡献度之和与其总实例个数的比率，即

$$FPR(c,o_i)=\frac{\sum Contri(FT(c),s_i)}{\left|FT(\{o_i\})\right|},(s_i\in FT(c)) \tag{9}$$

模式 $c=\{o_1,o_2,\cdots,o_k\}$ 的模糊参与度(FPI)是模式 c 中所有特征的模糊参与率最小值，即

$$FPI(c)=\min_{1\le i\le k}\{FPR(c,o_i)\} \tag{10}$$

给定一个用户自定义的最小模糊参与度阈值 min_*fprev*，对于任何一个模式 c，如果 $FPI(c)\ge$ min_*fprev*，则模式 c 是一个模糊频繁并置模式。

例 5：图 1 通过计算特征 A, B, C 在模式{A, B, C}中的模糊参与率分别为：$FPR(\{A,B,C\},A)=(0.8+0.8+0.5)/7=0.3$，$FPR(\{A,B,C\},B)=0.37$，$FPR(\{A,B,C\},C)=0.4$，所以模式 $FPI(\{A,B,C\})=\min(0.33,0.37,0.58)=0.3$。

定义 7 模糊损失率(FLR)与模糊损失度(FLI)：给定一个频繁并置模式 $c_k=\{o_1,o_2,\cdots,o_k\}$ $(k>2)$及其 c_k 的一个 $k-1$ 阶子模式 c_{k-1}，对于 c_{k-1} 及 c_k 中的任意一个特征 o_i ($o_i\in c_{k-1}$ & $o_i\in c_k$)，o_i 由 c_{k-1} 到 c_k 的模糊损失率定义为：

$$FLR(c_{k-1},c_k,o_i)=FPR(c_{k-1},o_i)-FPR(c_k,o_i) \tag{11}$$

模式由 c_{k-1} 到 c_k 的模糊损失度是 c_{k-1} 中的特征到 c_k 的模糊损失率的最小值

$$FLI(c_{k-1},c_k)=\min\{FLR(c_{k-1},c_k,o_i)\},o_i\in c_{k-1} \tag{12}$$

模式 c_{k-1} 到 c_k 的模糊损失度 $FLI(c_{k-1},c_k)$ 表示模式 c_{k-1} 中不与特征 $o_k(o_k=c_k-c_{k-1})$ 的实例满足邻近度阈值 α 的团实例的数量，当特征 $o_k(o_k=c_k-c_{k-1})$ 加入到模式 c_{k-1} 中形成高阶模式 c_k 时，模式 c_{k-1} 所损失的实例贡献度；损失的实例贡献度越高，特征 o_k 对模式 c_{k-1} 的影响程度越低，特征 o_k 对模式 c_{k-1} 中特

征的主导性越差。由此，给出特征影响度的定义。

定义 8 模糊影响度(FII)：给定一个频繁 co-location 模式 $c_k=\{o_1,o_2,\cdots,o_k\}$ $(k>2)$，特征 $o_i\left(o_i\in c_k,1\le i\le k\right)$，$c_k$的一个 $k-1$ 阶子模式 $c_{k-1}\left(o_i\notin c_{k-1}\right)$，则 o_i 在模式 c_k 中的模糊影响度定义为 1 减模式 c_{k-1} 到 c_k 模糊损失度，即

$$FII\left(c_k,o_i\right)=1-FLI\left(c_{k-1},c_k\right) \tag{13}$$

其中 $o_i=c_k-c_{k-1}$，特征模糊影响度 $FII\left(c_k,o_i\right)$ 代表了特征 o_i 对频繁模式 c_k 中其它特征的影响，$FII\left(c_k,o_i\right)$ 越大，越多 c_k 中的特征实例被特征 o_i 的实例主导。

例 6：给定 $\alpha=0.1$，则特征 A 从{A, B}到{A, B, C}的模糊损失率为：$FLR\left(\{\mathrm{A,B}\},\{\mathrm{A,B,C}\},\mathrm{A}\right)=0.1$，$FLR\left(\{\mathrm{A,B}\},\{\mathrm{A,B,C}\},\mathrm{B}\right)=0.03$，则{A, B}到{A, B, C}的模糊损失度为：$FLI\left(\{\mathrm{A,B}\},\{\mathrm{A,B,C}\}\right)=0.03$。同样的 $FLI\left(\{\mathrm{A,C}\},\{\mathrm{A,B,C}\}\right)=0.5$，$FLI\left(\{\mathrm{B,C}\},\{\mathrm{A,B,C}\}\right)=0.225$。则在频繁 co-location 模式{A, B, C}中，特征 A, B, C 对模式的模糊影响度分别为：$FII\left(\{\mathrm{A,B,C}\},\mathrm{A}\right)=1-0.225=0.775$，$FII\left(\{\mathrm{A,B,C}\},\mathrm{B}\right)=1-0.5=0.5$，$FII\left(\{\mathrm{A,B,C}\},\mathrm{C}\right)=1-0.03=0.97$。

定义 9 模糊影响比(FIR)：给定一个 k $(k>2)$阶频繁 co-location 模式 $c_k=\{o_1,o_2,\cdots,o_k\}$，特征 o_i，$o_j\left(o_i,o_j\in c_k\right)$满足 $FII(c_k, o_i)\ge FII(c_k, o_j)$，则 o_i 对 o_j 的模糊影响比定义为 1 减去两个特征的模糊影响度的比值，即

$$FIR\left(o_i,o_j\right)=1-\frac{FII\left(c_k,o_j\right)}{FII\left(c_k,o_i\right)} \tag{14}$$

传统主导特征挖掘算法是计算模式中两两特征差异度并设置差异度阈值来度量特征间的差异，但是本文通过模糊实例间的邻近关系，计算特征的差异度值很小，不能有效且明显的看出特征间的影响差异，所以提出了模糊影响比将特征间影响度的差异关系更为明显的表示出来。即模糊影响比指数更好地度量了模式中特征间的影响差异，模糊影响比指数 $FIR(o_i, o_j)$越大，特征 o_i 对特征 o_j 主导性越强。

例 7：在图 1 中模式{A, B, C}中特征 C 对特征 A 的影响比 $FIR\left(\mathrm{C,A}\right)=1-0.775/0.97=0.2$。特征 C 对特征 B 的影响比 $FIR\left(\mathrm{C,B}\right)=1-0.5/0.97=0.49$。

定义 10 主导特征和含主导特征的并置模式：给定一个 k $(k>2)$阶频繁并置模式 $c_k=\{o_1,o_2,\cdots,o_k\}$，最小模糊参与度阈值 min_*fprev* 和模糊影响比阈值 min_*fir*，对于 $o_i, o_j\in c_k$ 如果同时满足以下条件：1) $FII(c_k, o_i)\ge FII(c_k, o_j)$, 2) $FIR(o_i, o_j)\ge$ min_*fir*，那么在频繁并置模式中称特征 o_i 主导特征 o_j，且 o_i 是模式 c_k 的一个主导特征。对于并置模式 c_k 如果满足 $FPI(c_k)\ge$ min_*fprev*，则 c_k 是一个含主导特征并置模式。

例 8：给定模糊影响比阈值 min_*fir* = 0.1，通过计算模式{A, B, C}是含主导特征的并置模式，主导特征为 C，可以计算出特征 C 主导特征 A 和 B 的并置。

引理 1 反单调性。模糊参与率和模糊参与度随着模式阶的增大而单调非递增。

证明：设有 k 阶并置模式 $c_k=\{o_1,o_2,\cdots,o_k\}$ 和 $k+1$ 阶模式 $c_{k+1}=c_k+o_{k+1}$，对于 $o_i\in c_k$，如果 o_i 的某个实例含在 c_{k+1} 的模糊行实例中，那么该实例也一定包含在 c_k 的模糊行实例中，反之不然。由于实例的模糊行实例贡献度是指与其它实例间的最小邻近度，实例的模式贡献度是取某个实例重复出现在多个行实例时贡献度的最大值并参与特征的模糊参与率计算，因此 o_i 的某个实例对 o_i 在 c_{k+1} 中的计算模糊参与率时贡献度值一定不大于该实例对 o_i 在模式 c_k 中参与模糊参与率计算的贡献值。因此，模糊参与率随着模式阶的增大而单调非递增。因为模糊参与度是取模式中所有特征模糊参与率的最小值，所以模糊参与率也随着模式阶的增大而单调非递增。即证。

定理 1 向下闭合性。在基于模糊邻近关系(FNR)的并置模式挖掘中，如果一个并置模式 c 是频繁的，

则它的所有子模式 $c' \subset c$ 都是频繁的；相反的，如果一个并置模式 c 是非频繁的，则它的所有超模式 $c' \supset c$ 也都是非频繁的。

证明：由引理 1 可知，并置模式 c 的模糊参与度不大于其所有子模式的模糊参与度，不小于其超模式的模糊参与度，所以如果并置模式 c 是频繁的，则其所有子模式一定都是频繁的；如果并置模式 c 是非频繁的，则其所有超集也一定都是非频繁的。即证。

3. 挖掘算法

本节给出了挖掘含主导特征的空间并置模式的基本算法 DFMAFPR 和优化算法 DFMAFPR-Improved。

3.1. DFMAFPR 算法

根据用户给定的邻近度阈值 α 和最小影响比阈值 min_*fir*，首先挖掘出所有模糊频繁的并置模式，计算特征的模糊损失率和模式的模糊损失度，最后通过特征间的模糊影响比挖掘出含有主导特征的频繁并置模式。具体见算法 1：

算法 1. DFMAFPR

输入：

O：空间特征集，S：空间实例集，μ：模糊邻近关系的隶属度函数，α：邻近度阈值，min_*fprev*：最小模糊参与度阈值，min_*fir*：最小特征模糊影响比阈值

变量：

K：co-location 模式的阶，FSN：模糊星型邻居集，C_k：k 阶候选 co-location 模式，FPR_c：k 阶 co-location 模式 c 的模糊参与率集，P_k：k 阶频繁 co-location 模式集，P：频繁 co-location 模式集

输出：

含主导特征的 co-location 频繁模式集 DFCP_*set* 及所有 DFCP 的主导特征集

步骤：

1) FNR = get_fuzzy_neighbor_relationship(S, μ);
2) FSN = get_star_neighbor(O, S,FNR$_\alpha$);
3) $P_1 = O$, $k = 2$, DFCP = ∅;
4) while ($P_{k-1} \neq \varnothing$) do
5) C_k = gen_candidate_fuzzy_co-location(k, P_{k-1});
6) for each $c \in C_k$ do
7) if calculate FPI(c) ≥ min_*fprev* do
8) for each $p \in P_{k-1}(c)$ and FPR_c do
9) FLI(p, c) = calculate_FLI(FPR(p), FPR(c))
10) FII_set(c) ← {1-FLI(p, c), c-p}
11) end do
12) for each o_i, $o_j \in c$ do
13) if FIR(o_i,o_j) ≥ min_*fir* do
14) DF_*set*(c) ← o_i;
15) end do
16) if DF_*set*(c) ≠ ∅ do
17) DFCP ← {c, DF_*set*(c)};
18) end do
19) end do
20) end do
21) $k = k + 1$
22) end do
23) Output DFCP

第 1、2 行根据给定模糊邻近关系的隶属函数，使用网格划分技术，计算空间数据集的模糊邻近关系，获取满足邻近度阈值 α 的模糊邻近对，并生成模糊星型邻居集。第 3~5 行生成 k 阶候选模式集，第 6~7 行计算候选模式的模糊参与度。对于满足模糊参与度阈值的模式，第 8~11 行计算该模式的 $k-1$ 阶子模式集合计算模糊损失度并得到每个特征的模糊影响度。第 12~15 行取出模式中所有特征的特征模糊影响度，并计算模式中两两特征模糊影响度的特征模糊影响比是否满足特征模糊影响比阈值 min_*fir*，将满足阈值的特征 o_i 放入主导特征集合 DF_*set*(c)。第 17~19 行存储 DFCP 及其主导特征集集合中。第 4~22 行被重复执行，用于逐阶输出所有 DFCP 及其主导特征集。

算法 1 分析：DFMAFPR 算法的时间复杂度分析可以分为四个部分，由于 DFMAFPR 算法是由 Joinless 算法改进得到，所以前三个部分和 Joinless 相似，第一部分复杂度为 $T_{f\_star\_neib}$，区别于 Joinless 进一步将实例距离模糊化后转化为邻近度。第二部分复杂度是 $T(2)$生成二阶频繁模式消耗的时间，第三部分复杂度是 $\sum_{k=2}^{h} T(k)$，表示生成高阶频繁模式消耗的时间，其中 h 表示迭代时 co-location 模式的最高阶数。第四部分复杂度是 T_{DF_set}，表示在频繁模式中挖掘主导特征消耗的时间。由于 DFMAFPR 算法考虑了实例间的邻近程度，在生成相同数量的候选 co-location 模式情况下，时间复杂度比 Joinless 算法更高一些。相比于同样改进 Joinless 算法后的主导特征模式挖掘算法 ADFSPTCM [21]，本文所提出的 DFMAFPR 算法复杂度也会更高一些，当相关参数一致的情况下 DFMAFPR 算法挖掘出的主导特征模式比 ADFSPTCM 算法更有价值。

3.2. 优化算法 DFMAFPR-Improved

该算法是在算法 1 基础上进一步优化主导特征的挖掘过程，在筛选出的频繁模式中计算最大特征影响度和最小特征影响度后，计算最大特征影响比，当最大特征影响比满足给定阈值时，保留该频繁模式进行主导特征挖掘计算，可以在计算特征影响比之前对频繁模式进行剪枝，进一步提高挖掘效率。

算法 2. DFMAFPR-Improved

输入：
FII_set(c)：频繁并置模式及各特征影响度，min_*fir*：最小特征模糊影响比阈值
输出：
最小和最大特征模糊影响度之比满足模糊影响比阈值的频繁模式 P
步骤：
1) if FII_set(c) $\neq \varnothing$ do
2) $o_{\min}$ = argmin{FII_set(c)}, $o_{\max}$ = argmax{FII_set(c)}
3) if max_FIR($o_{\min}$, $o_{\max}$) $\geq$ min_*fir* do
4) $P = c$
5) end do
6) end do
7) Output P

第 1 行判断初始频繁并置模式集合非空；第 2 行取模式中最小特征模糊影响度和最大特征模糊影响度；第 3 行计算最大特征模糊影响比并验证是否满足最大特征模糊影响比阈值；第 4 行将满足阈值的频繁并置模式放入集合。

算法 2 分析：根据相关定义，在含主导特征的并置模式挖掘过程中，对于一个模糊频繁并置模式 c_k 需要计算所有特征两两之间的特征影响比，直至确认该模式中没有任何一对特征满足特征影响比阈值，

才认为不含主导特征。DFMAFPR-Improved 为了优化挖掘过程，提取模式中最大特征模糊影响度和最小特征模糊影响度来判断模式中是否含有主导特征，通过 k 次比较完成。即这一阶段的计算复杂度仅为模式长度 k，该过程大大加速了挖掘速度。所以 DFMAFPR-Improved 的时间复杂度在第四部分大大减少了计算时间。

4. 实验结果与分析

本节基于合成数据集和真实数据集，对于我们所提出的基于模糊邻近关系的含主导特征的并置模式挖掘算法(DFMAFPR)和优化算法(DFMAFPR-Improved)做实验评价。主要目的是：1) 在合成数据集上，我们评估不同实验参数对算法效率的影响以及算法的可伸缩性；2) 在合成数据集和真实数据集上，对比 DFMAFPR 和 DFMAFPR-Improved 算法和传统主导特征模式挖掘算法(ADFCPM) [21]挖掘结果的差异；3) 在真实数据集上，给出含主导特征的并置模式挖掘结果实例，进一步说明基于模糊邻近关系的含主导特征并置模式挖掘算法在实际应用中的意义。

运行环境：所有算法采用 python 语言实现，并运行于具有 AMD Ryzen 7 1700X Eight-Core Processor 3.40 GHz 处理器，16 GB 内存、Windows 10 及 pycharm 2020 的 PC 机上。参数设置：所提主导特征模式挖掘算法 ADFSPTCM、DFMAFPR-Improved 在各个数据集上的实验参数默认设置如表 2 所示。

4.1. 实验数据集

实验数据集：实验采用了 3 个合成数据集和 2 个真实数据集，数据集相关信息如表 1 所示，其中，Plant-Data 是一个包含 31 种植物类型(特征)共 356 棵植物(实例)的“三江并流”区域珍稀植物数据集。北京 POI 是一个包含 16 种类型(特征)共 2305 个 POI (实例)的北京市 POI 数据集。合成数据集采用与文献[21]提出的数据生成器类似的方法根据泊松分布函数分别在 500 × 500、1000 × 1000 的范围内生成合成数据。

Table 1. Experimental data set parameters
表 1. 实验数据集参数

数据集	特征集	实例数	范围
Plantdata	32	335	8000 × 13,000
Beijing-POI	16	23,025	22,000 × 14,000
合成数据集 1	10	10,000	500 × 500
合成数据集 2	10	10,000	1000 × 1000
合成数据集 3	25	50,000	1000 × 1000

Table 2. Default parameter description
表 2. 实验默认说明

数据集(Dataset)	距离阈值(d/d_1)	模糊参与度阈值(min_*fprev*)	影响比阈值(min_*fir*)
Plantdata	5000	0.3	0.3
Beijing-POI	50	0.3	0.3
合成数据集 1	20	0.3	0.3
合成数据集 2	20	0.3	0.3
合成数据集 3	20	0.3	0.3

4.2. 不同参数对算法性能的影响

本文所提算法所用距离阈值 $d_2 = 80$，我们在 3 个不同规模的合成数据集上分析不同参数设置对所提主导特征模式挖掘算法 DFMAFPR、优化算法 DFMAFPR-Improved 进行比较。目的是观察在不同实例数量和特征数量下，参数变化对算法造成的不同影响。

4.2.1. 距离阈值 *d* 对算法性能的影响

图 3~5 分别显示两个算法在 d_1 取 20、30、40 和 50 四个不同的距离阈值时在 3 个合成数据集上的性能表现，其它参数取表 2 中默认值。在每个数据集上，随着距离阈值的增大，算法运行时间逐渐增加，并且随着数据集的规模增大，运行时间也逐渐增加。距离阈值较大时，对算法性能的影响较为明显，这说明算法性能受到数据稠密性的影响。合成数据集 1 比合成数据集 2 分布更稠密，邻近度阈值的影响较明显，并且运行时间也相对较长。合成数据 3 的特征数量和实例数量最多，所以运行时间最长。算法 DFMAFPR-Improved 表现比基础算法 DFMAFPR 更好，是因为在计算主导特征时有效的对模式进行剪枝。

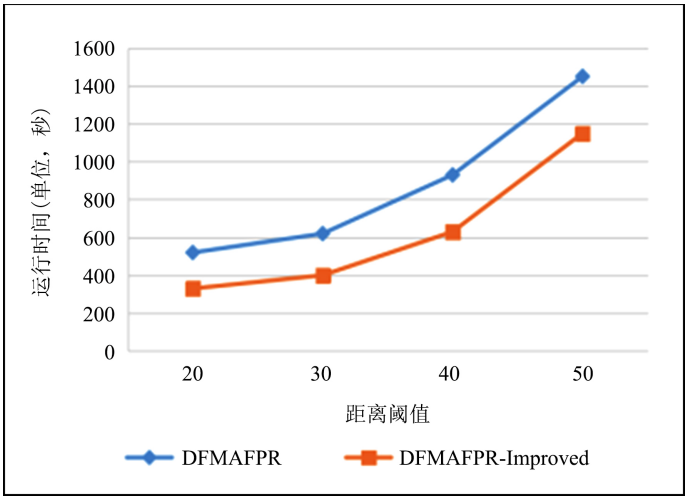

Figure 3. Performance comparison of different distance thresholds (Synthetic data 1)
图 3. 不同距离阈值性能比较(合成数据集 1)

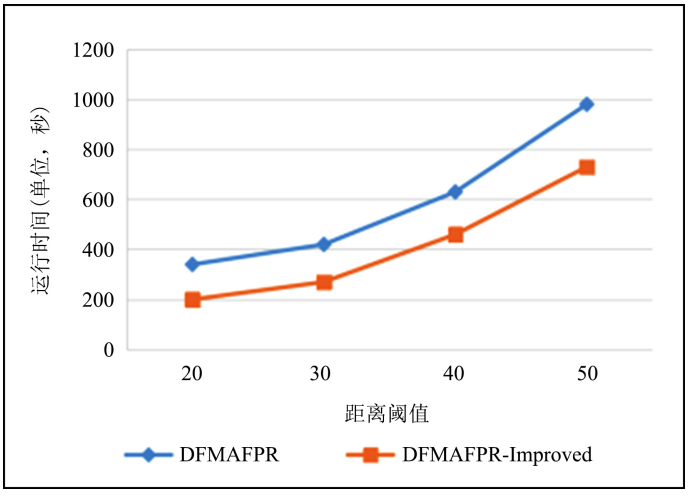

Figure 4. Performance comparison of different distance thresholds (Synthetic data 2)
图 4. 不同距离阈值性能比较(合成数据集 2)

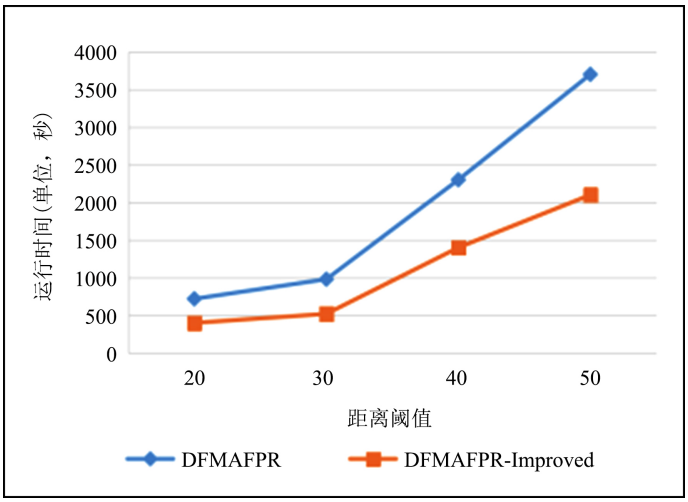

Figure 5. Performance comparison of different distance thresholds (Synthetic data 3)
图 5. 不同距离阈值性能比较(合成数据集 3)

4.2.2. (模糊)参与度阈值对算法性能的影响

图 6~8 分别显示在(模糊)参与度阈值分别取 0.3，0.4，0.5 和 0.6 时两个算法在 3 个合成数据集上的性能表现，其它参数取表 2 中默认值。随着数据规模的增加和(模糊)参与度阈值的增加，运行时间逐渐减少。合成数据集 1 比合成数据集 2 分布更稠密，同一范围内实例数量增加，满足成团的实例增多，对(模糊)表实例的运算耗费影响了算法性能，(模糊)参与度阈值变化对算法效率的影响较为明显，其中合成数据集 3 的运行时间最长。

4.2.3. 特征影响比阈值对算法性能的影响

图 9~11 分别显示 DFMAFPR 算法、DFMAFPR-Improved 算法在 3 个合成数据集上的性能，其它参数取表 2 中默认值。特征模糊影响比阈值 min_*fir* 分别取 0.3、0.4、0.5 和 0.6，随着特征模糊影响比阈值的变化，算法运行时间逐渐减。因为随着阈值的升高，需要计算的频繁模式表实例减少，阈值的变化对于稠密数据集上算法性能的影响更为明显。

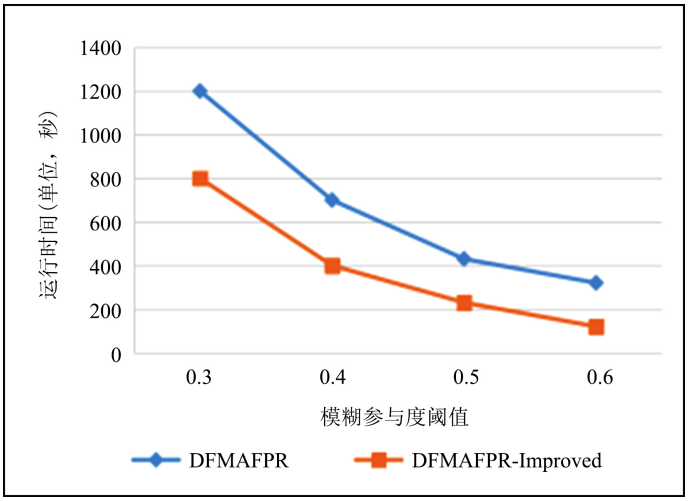

Figure 6. Performance comparison of different engagement thresholds (Synthetic data 1)
图 6. 不同参与度阈值性能比较(合成数据集 1)

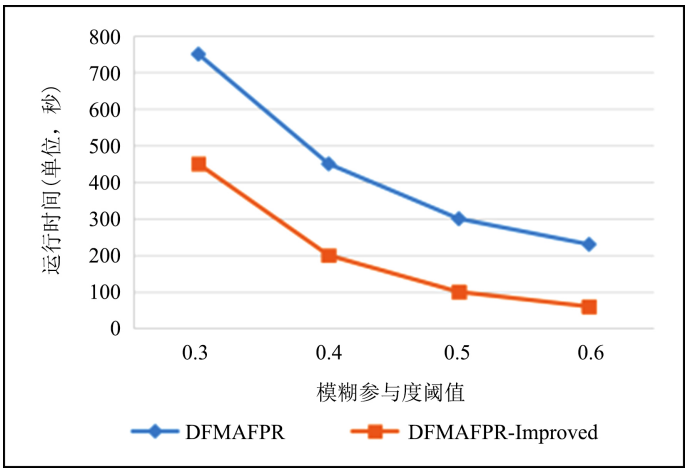

Figure 7. Performance comparison of different engagement thresholds (Synthetic data 2)
图 7. 不同参与度阈值性能比较(合成数据集 2)

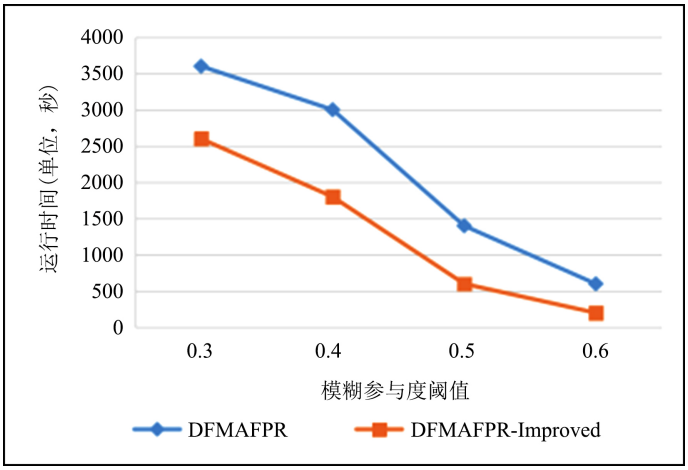

Figure 8. Performance comparison of different engagement thresholds (Synthetic data 3)
图 8. 不同参与度阈值性能比较(合成数据集 3)

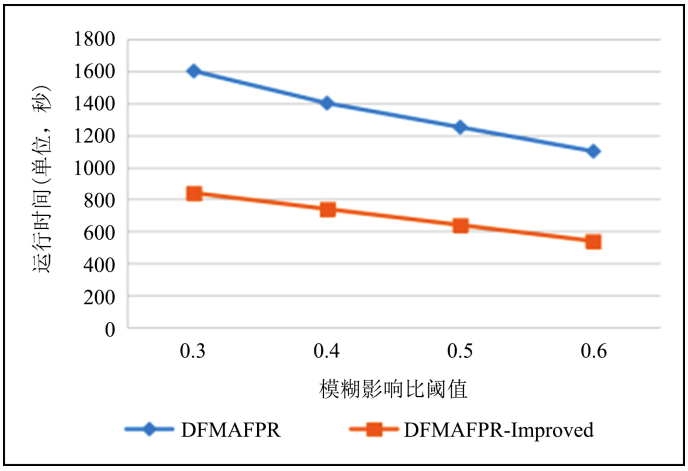

Figure 9. Performance comparison of different impact ratio thresholds (Synthetic data 1)
图 9. 不同影响比阈值性能比较(合成数据集 1)

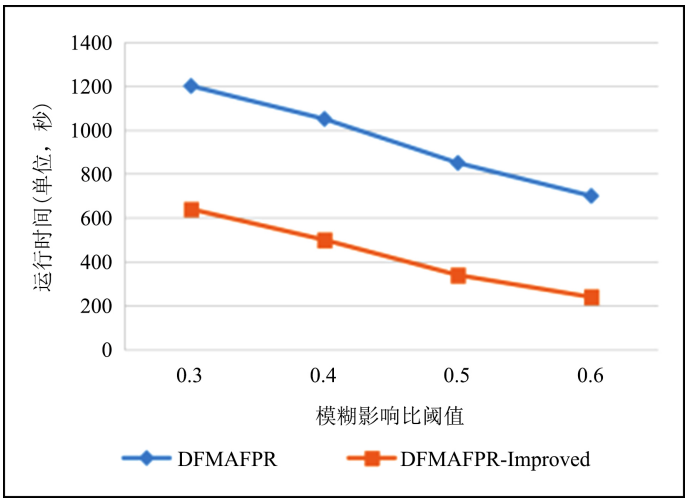

Figure 10. Performance comparison of different impact ratio thresholds (Synthetic data 2)
图 10. 不同影响比阈值性能比较(合成数据集 2)

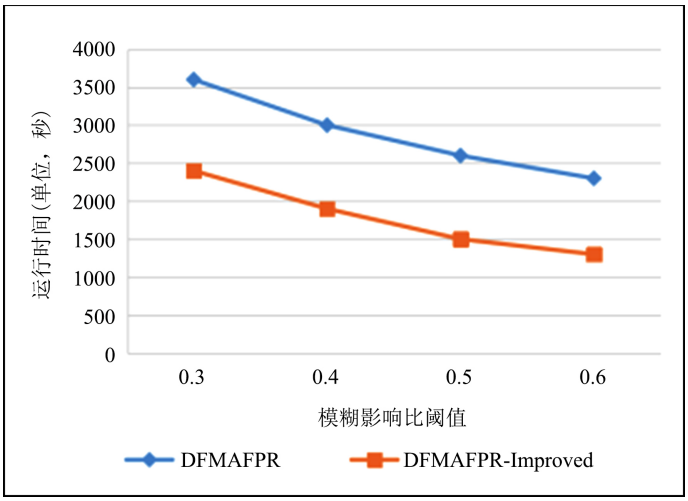

Figure 11. Performance comparison of different impact ratio thresholds (Synthetic data 3)
图 11. 不同影响比阈值性能比较(合成数据集 3)

通过本文所提两种算法和传统主导特征模式挖掘算法在 2 个合成数据集上的比较，优化后的 DFMAFPR-Improved 算法在基础算法上消耗时间更少，在挖掘主导特征过程中减少大量的计算。

4.3. DFMAFPR 算法的挖掘结果分析

我们在 2 个真实数据集上，比较分析所提基于模糊邻近关系主导特征模式挖掘算法 DFMAFPR 与传统频繁模式挖掘算法 JoinLess、传统含主导特征模式挖掘算法 AMDFCP 的挖掘结果。

4.3.1. 在“三江并流”区域珍稀植物数据集上的结果比较

在“三江并流”区域珍稀植物数据集上，三个算法在不同(模糊)参与度阈值下挖掘得到的模式数量如图 12 所示，其中 DFMAFPR 算法两个距离阈值 d_1 = 5000 和 d_2 = 15000，Joinless 算法和 AMDFCP 算法所用距离阈值 d = 5000，其它参数取表 2 中默认值。从图中可以看出 DFMAFPR 挖掘的主导特征模式数量大约为 Joinless 算法挖掘的频繁模式数量的 50%，这是因为 DFMAFPR 有效去除了不含主导特征的频

繁模式。我们提出的算法挖出的含主导特征模式比 AMDFCP 算法的含主导特征模式数量多，是因为 AMDFCP 算法是基于单一邻近阈值判断实例间邻近关系，实例在形成团关系时会造成邻近关系缺失。

三个算法在不同距离阈值 d 下挖掘得到的模式数量如图 13 所示。随着距离阈值 d 的增大，三个算法挖掘到的频繁模式的数量逐渐增加，其中 Joinless 模式数量增加趋势较为明显，这是因为在数据分布密度较高时，随着同一邻域内的实例数量和特征数量类型增多，候选模式的数量和团实例的数量都急剧增加；我们还可以看到三种主导特征模式挖掘算法产生的模式数量远远小于 Joinless 算法产生的模式数量，但是本文所提算法挖掘出的主导特征模式数量都比 AMDFCP 算法多。

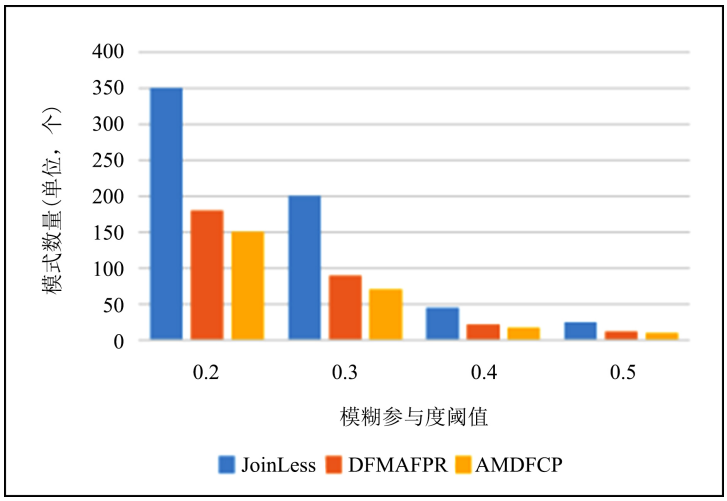

Figure 12. The number of patterns with different min_*fprev* on Plant-Data
图 12. 植物数据集上不同(模糊)参与度阈值下的模式数量

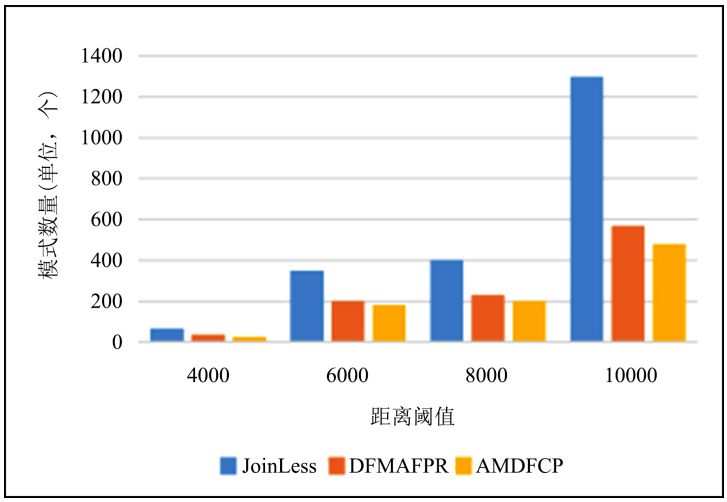

Figure 13. The number of patterns with different d on Plant-Data
图 13. 植物数据集上不同距离阈值下的模式数量

4.3.2. 在北京 POI 数据集上的结果比较

在北京 POI 数据集上，三个算法在不同(模糊)参与度阈值、距离阈值下挖掘得到的模式数量分别如图 14 和图 15 所示，其中 DFMAFPR 算法两个距离阈值取 $d_1 = 50$ 和 $d_2 = 150$，Join Less 算法和 AMDFCP 算法所用距离阈值 $d = 50$，其它参数取表 2 中默认值。

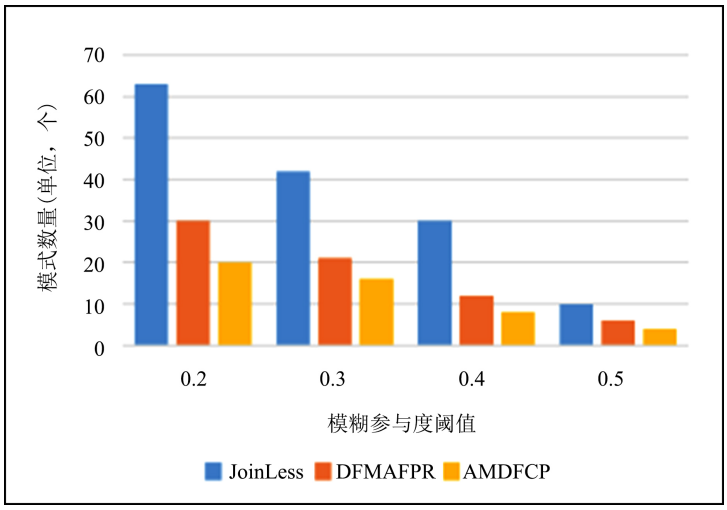

Figure 14. The number of patterns with different min_*fprev* on Beijing POI
图 14. 北京 POI 数据集上不同(模糊)参与度阈值下的模式数量

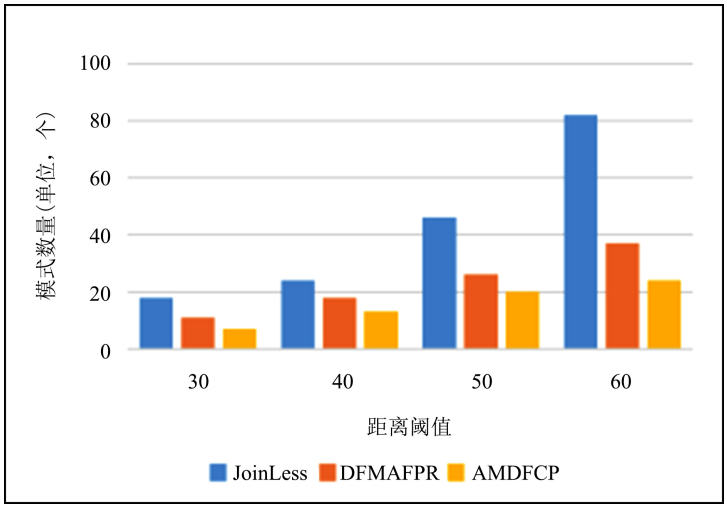

Figure 15. The number of patterns with different *d* on Beijing POI
图 15. 北京 POI 数据集上不同距离阈值 *d* 下的模式数量

通过在两个真实数据集中的比较，本文所提出的算法 DFMAFPR 虽然受距离阈值 d_1 影响，但与传统挖掘算法 AMDFCP 相比，DFMAFPR 挖掘出主导特征模式数量更多。真实数据集 1 中包含 32 个特征，但仅有 335 个实例，在挖掘主导特征时(模糊)表实例中的实例数量变化将导致特征(模糊)参与率的计算。

4.4. 真实数据集上的挖掘结果及应用实例分析

为了验证本文所提主导特征模式的实用性，我们对本文所提含主导特征模式挖掘算法和传统 AMDFCP 算法在 2 个真实数据集上挖掘得到的含主导特征模式进行实例分析。表 3 和表 4 分别列出了 DFMAFPR 算法和 AMDFCP 算法在植物数据集和北京 POI 数据集上的一些三阶模式。从表中我们可以看出在植物数据集挖掘出的模式{川八角莲，珙桐，冬虫夏草}、{松茸，丽江雪胆，长苞冷杉}和{松茸，三尖杉，穿心延子薷}是 DFMAFPR 和 AMDFCP 共有的主导特征模式，其中主导特征相同，而{松茸，三尖杉，长苞冷杉}和{云南榧木，三尖杉，红豆杉}是 AMDFCP 算法挖掘不到的主导特征模式。在北京 POI

数据集挖掘出{酒店，停车场，服装店}、{酒店，中餐馆，停车场}、{西餐厅，停车场，公交站}是 DFMAFPR 和 AMDFCP 共有的主导特征模式，其中主导特征相同，而{花园，停车场，咖啡厅}、{中餐厅，咖啡厅，招待所}是 AMDFCP 算法挖掘不到的主导特征模式。

Table 3. The mining results DFMAFPR and AMDFCP on "Three Parallel Rivers"
表 3. DFMAFPR 和 AMDFCP 在三江并流植物数据上挖掘的结果比较

含主导特征的并置模式	DFMAFP 算法		AMDFCP 算法	
	FPR	FIR	FR	FD
{**川八角莲***，**珙桐***，冬虫夏草}	0.3	0.3	0.3	0.2
{松茸，丽江雪胆，**长苞冷杉***}	0.3	0.3	0.3	0.2
{松茸，**三尖杉***，**穿心延子薰***}	0.3	0.3	0.3	0.2
{松茸，**三尖杉***，长苞冷杉}	0.3	0.3	-	-
{**云南榧木***，**三尖杉***，红豆杉}	0.3	0.3	-	-

注：*代表主导特征；-代表没有参数可挖掘到。

Table 4. The mining results DFMAFPR and AMDFCP on "Beijing POI"
表 4. DFMAFPR 和 AMDFCP 在北京 POI 数据上挖掘的结果比较

含主导特征的并置模式	DFMAFP 算法		AMDFCP 算法	
	FPR	FIR	FR	FD
{**酒店***，**停车场***，服装店}	0.4	0.4	0.4	0.4
{**酒店***，**中餐馆***，停车场}	0.4	0.4	0.4	0.4
{**西餐厅***，停车场，公交站}	0.4	0.4	0.4	0.4
{**花园***，停车场，**咖啡厅***}	0.4	0.4	-	-
{中餐厅，**咖啡厅***，**招待所***}	0.4	0.4	-	-

注：*代表主导特征；-代表没有参数可挖掘到。

首先，本文所提算法和 AMDFCP 算法在挖掘主导特征模式时都是从三阶频繁模式开始，但是 AMDFCP 算法局限于单一的邻近阈值，超出邻近阈值的实例对就判定为不邻近，这种单一截断的邻近关系判断会造成邻近关系的损失，并且 AMDFCP 算法在挖掘频繁模式时对于频繁性阈值非常敏感，当两两实例对于单一距离阈值判为不邻近时，就不参与实例特征参与率计算，但是 DFMAFPR 和 DFMAFPR-Improved 算法是使用模糊数学的方法去计算两两实例的邻近程度，首先在挖掘主导特征并置模式时计算实例对模式的模糊贡献度，其次通过模糊参与率和模糊参与度挖掘出更有参考价值的频繁并置模式，进而发掘模式中两两特征间的关系并挖掘出含有主导特征的并置模式。

在表 3 和表 4 中，通过比较 2 个真实数据集挖掘出来的结果发现，特征数量和实例数量较多的数据集能发现更多含有主导特征的模式，更好的揭示模式中特征的主导关系。通过实验分析，虽然 DFMAFPR 挖掘效果受两个距离阈值 d_1 和 d_2 影响，但与 AMDFCP 算法相比挖掘出的模式数量更多，本文所提算法挖掘出的主导特征模式更有价值和决策性，当 d_1 和 d_2 取值相近时两个算法挖掘模式数量几乎相同。通过

本文所提两种算法和传统主导特征模式挖掘算法在 3 个合成数据集和 2 个真实数据集上的比较，优化后的 DFMAFPR-Improved 算法比基础算法消耗时间更少，由于算法在计算实例间的邻近度和模式中模糊参与率、模糊参与度时更耗费时间，所以 DFMAFPR 算法与传统 AMDFCP 算法比较时间效率相对较差，但是挖掘出的模式会更有价值，可以更好的应用在实际生活中。

5. 总结

主导关系体现的是中心事物对周边事物的吸引力或者周边事物对中心事物的依赖性，本文基于模糊空间实例邻近关系，研究空间频繁并置模式的主导特征挖掘，以更好地揭示空间特征间的主导关系。首先，本文在模糊邻近关系的基础上，给出实例对模糊行实例、模式的贡献、模糊参与率和模糊参与度等相关定义，并通过特征影响比指标度量特征的主导性。然后，提出了有效的含主导特征模式挖掘算法。最后，通过在合成数据集和真实数据集上的大量实验验证了本文所提算法能够挖掘出更多更有价值的含主导特征的并置模式。在未来的研究中，我们将设计高效的剪枝策略和挖掘方法，进一步提高算法的挖掘效率。

基金项目

国家自然科学基金项目(No.61966036, No.61662086)；云南省创新团队项目(No.2018HC019)。

参考文献

[1] Akbari, M., Samadzadegan, F. and Weibel, R. (2015) A Generic Regional Spatio-Temporal Co-Occurrence Pattern Mining Model: A Case Study for Air Pollution. *Journal of Geographical Systems*, **17**, 249-274. https://doi.org/10.1007/s10109-015-0216-4

[2] Yu, W., Ai, T., He, Y., *et al*. (2017) Spatial Co-Location Pattern Mining of Facility Points-of-Interest Improved by Network Neighborhood and Distance Decay Effects. *International Journal of Geographical Information Science*, **31**, 280-296. https://doi.org/10.1080/13658816.2016.1194423

[3] An, S., Yang, H.Q., Wang, J., *et al*. (2016) Mining Urban Recurrent Congestion Evolution Patterns from GPS Equipped Vehicle Mobility Data. *Information Sciences*, **373**, 515-526. https://doi.org/10.1016/j.ins.2016.06.033

[4] Wang, L.Z. and Chen, H.M. (2014) Spatial Pattern Mining Theory and Methods. Science Press, Beijing, 2-4.

[5] Fang, Y., Wang, L.Z., Wang, X.X., *et al*. (2017) Mining Co-Location Patterns with Dominant Features. In: *International Conference on Web Information Systems Engineering*, Springer, Cham, 183-198. https://doi.org/10.1007/978-3-319-68783-4_13

[6] Wang, L.Z., Bao, X.G., Zhou, L.H., *et al*. (2017) Maximal Sub Prevalent Co-Location Patterns and Efficient Mining Algorithms. In: *International Conference on Web Information Systems Engineering*, Springer, Cham, 199-214. https://doi.org/10.1007/978-3-319-68783-4_14

[7] Wang, L.Z., Bao, X.G., Zhou, L.H., *et al*. (2019) Mining Maximal Sub Prevalent Co-Location Patterns. *World Wide Web*, **22**, 1971-1997. https://doi.org/10.1007/s11280-018-0646-2

[8] Huang, Y., Shekhar, S. and Xiong, H. (2004) Discovering Co-Location Patterns from Spatial Data Sets: A General Approach. *IEEE Transaction s on Knowledge and Data Engineering*, **16**, 1472-1485. https://doi.org/10.1109/TKDE.2004.90

[9] Yoo, J.S., Shekhar, S., Smith, J., *et al*. (2004) A Partial Join Approach for Mining Co-Location Patterns. *Proceedings of the 12th Annual ACM International Workshop on Geographic Information Systems*, Arlington, 12-13 November 2004, 241-249. https://doi.org/10.1145/1032222.1032258

[10] Yoo, J.S., Shekhar, S. and Celik, M. (2005) A Join-Less Approach for Co-Location Pattern Mining: A Summary of Results. *Proceedings of the 5th IEEE International Conference on Data Mining* (*ICDM*), Houston, 27-30 November 2005, 813-816.

[11] Wang, L.Z., Bao, Y., Lu, J.L., *et al*. (2008) A New Join-Less Approach for Co-Location Pattern Mining. 2008 8*th IEEE International Conference on Computer and Information Technology*, Sydney, 8-11 July 2008, 197-202.

[12] Wang, L.Z., Bao, Y. and Lu, Z. (2009) Efficient Discovery of Spatial Co-Location Patterns Using the ICPI-Tree. *The Open Information Systems Journal*, **3**, 69-80. https://doi.org/10.2174/1874133900903020069

[13] Wang, L.Z., Zhou, L.H., Lu, J.L., *et al*. (2009) An Order-Clique-Based Approach for Mining Maximal Co-Locations. *Information Sciences*, **179**, 3370-3382. https://doi.org/10.1016/j.ins.2009.05.023

[14] Wang, L.Z., Chen, H.M., Zhao, L., *et al*. (2010) Efficiently Mining Co-Location Rules on Interval Data. In: *International Conference on Advanced Data Mining and Applications*, Springer, Berlin, 477-488. https://doi.org/10.1007/978-3-642-17316-5_45

[15] Lu, Y., Wang, L.Z. and Zhang, X.F. (2009) Mining Frequent Co-Location Patterns from Uncertain Data. *Journal of Frontiers of Computer Science and Technology*, **3**, 656-664.

[16] Wang, L.Z., Wu, P. and Chen, H.M. (2013) Finding Probabilistic Prevalent Co-Locations in Spatially Uncertain Data Sets. *IEEE Transactions on Knowledge and Data Engineering*, **25**, 790-804. https://doi.org/10.1109/TKDE.2011.256

[17] Huang, Y., Pei, J. and Xiong, H. (2006) Mining Co-Location Patterns with Rare Events from Spatial Data Sets. *Geoinformatica*, **10**, 239-260. https://doi.org/10.1007/s10707-006-9827-8

[18] Feng, L., Wang, L.Z. and Gao, S.J. (2012) A New Approach of Mining Co-Location Patterns in Spatial Datasets with Rare Features. *Journal of Nanjing University* (*Natural Sciences*), **48**, 99-107.

[19] Ouyang, Z.P., Wang, L.Z. and Chen, H.M. (2011) Mining Spatial Co-Location Patterns for Fuzzy Objects. *Chinese Journal of Computers*, **34**, 1947-1955. https://doi.org/10.3724/SP.J.1016.2011.01947

[20] Fang, Y., Wang, L.Z. and Hu, T. (2018) Spatial Co-Location Pattern Mining Based on Density Peaks Clustering and Fuzzy Theory. In: *Asia-Pacific Web* (*APWeb*) *and Web-Age Information Management* (*WAIM*) *Joint International Conference on Web and Big Data*, Springer, Cham, 298-305. https://doi.org/10.1007/978-3-319-96893-3_22

[21] Fang, Y., Wang, L.Z. and Zhou, L.H. (2016) Research on Mining Significant Co-Location Pattern with Key Features. *Data Acquisition and Processing*, **33**, 692-703.

[22] Ma, D., Chen, H.M., Wang, L.Z. and Xiao, Q. (2020) Dominant Feature Mining of Spatial Sub-Prevalent Co-Location Patterns. *Journal of Computer Applications*, **40**, 465-472.

[23] Lei, L., Wang, L.Z. and Xiao, Q. (2019) Study on Fuzzy Mining Technology in Spatial Co-Location Pattern Mining. *CEA*, **55**, 158-166.

[24] Wang, X.X., Wang, L.Z. and Wang, J.L. (2020) Mining Spatio-Temporal Co-Location Fuzzy Congestion Patterns from Traffic Datasets. *Journal of Tsinghua University* (*Science and Technology*), **60**, 683-692.

[25] Lei, L., Wang, L.Z. and Wang, X.X. (2019) Mining Spatial Co-Location Patterns by Fuzzy Technology. *Proceedings of the* 2019 *IEEE International Conference on Big Knowledge* (*ICBK*), Beijing, 10-11 November 2019, 129-136. https://doi.org/10.1109/ICBK.2019.00025