Hans Journal of Data Mining 数据挖掘, 2024, 14(2), 65-80
Published Online April 2024 in Hans. https://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2024.142006

Hans汉斯

基于Voronoi图和距离衰减效应的模糊实例空间并置模式挖掘算法

陈书健，芦俊丽

云南民族大学数学与计算机科学学院，云南 昆明

收稿日期：2024年3月1日；录用日期：2024年4月1日；发布日期：2024年4月9日

摘　要

空间并置模式挖掘用于发现一组空间特征，它们的实例在空间中频繁地相互邻近。传统的空间并置模式挖掘过程中，将空间实例抽象成点对象，每个实例对应一个确定位置。然而，规模较大的空间实例有多个重要位置点(如医院、公园入口)，其空间位置因对其入口的认知不同而存在差异，具有模糊性。对于这些模糊实例，本文考虑其重要位置点对该实例规模的贡献，重新定义实例间的邻近度。此外，传统的并置模式挖掘方法忽略了特征实例的空间分布密度以及邻近实例间的邻近程度，采用静态的距离阈值来识别邻近实例。本文考虑特征的分布密度，用Voronoi图自适应提取不同特征的邻近实例，结合邻近实例的距离衰减函数，更加科学地描述实例间的邻近度。提出一种同时考虑模糊实例规模和距离衰减效应的空间并置模式挖掘方法，为实现快速挖掘，设计了极大团和哈希表搜索参与实例的挖掘框架。在真实数据集和合成数据集上进行实验，验证本文的算法可以发现传统空间并置模式挖掘方法所忽略的有意义模式。

关键词

空间并置模式，模糊实例，距离衰减效应，Voronoi图

Fuzzy Instance Space Co-Location Pattern Mining Algorithm Based on Voronoi Graph and Distance Attenuation Effect

Shujian Chen, Junli Lu

College of Mathematics and Computer Science, Yunnan Minzu University, Kunming Yunnan

Received: Mar. 1st, 2024; accepted: Apr. 1st, 2024; published: Apr. 9th, 2024

文章引用: 陈书健, 芦俊丽. 基于 Voronoi 图和距离衰减效应的模糊实例空间并置模式挖掘算法[J]. 数据挖掘, 2024, 14(2): 65-80. DOI: 10.12677/hjdm.2024.142006

Abstract

Space co-location pattern mining is used to discover a set of spatial features whose instances are frequently adjacent to each other in space. In the process of traditional space co-location pattern mining, spatial instances are abstracted into point objects, and each instance corresponds to a definite location. However, large spatial examples have multiple important location points (such as hospital, and park entrance), and their spatial locations are different due to different cognition of their entrances, which is fuzzy. For these fuzzy instances, this paper considers the contribution of important location points to the instance scale, and redefines the proximity between the instances. In addition, the traditional collocation pattern mining method ignores the spatial distribution density of feature instances and the proximity degree between neighboring instances, and uses static distance threshold to identify neighboring instances. In this paper, considering the distribution density of features, Voronoi diagram is used to adaptively extract adjacent instances with different features, and the proximity of adjacent instances is described more scientifically by combining the distance decay function of adjacent instances. In this paper, a space co-location pattern mining method is proposed, which takes into account both the fuzzy instance size and distance attenuation effects. In order to realize fast mining, a mining framework for the participating instances of maximal clique and hash table search is designed. Experiments are carried out on real data sets and synthetic data sets to verify that the proposed algorithm can find meaningful patterns ignored by traditional space co-location pattern mining methods.

Keywords

Space Co-Location Pattern, Fuzzy Instance, Distance Attenuation Effect, Voronoi Diagram

Copyright © 2024 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

空间并置模式挖掘是空间数据挖掘的一个重要分支。从空间数据集中发现频繁空间并置模式就是发现一组空间特征，其实例频繁地相互邻近。例如，在一个城市的兴趣点(POI)数据集中，通过使用频繁空间并置模式挖掘技术，我们发现银行、超市和汽车站这 3 个特征的实例经常在邻近区域内一起出现，我们就称{银行、超市、汽车站}是一个频繁空间并置模式。频繁空间并置模式挖掘技术可以发现空间数据集中特征之间的内在联系，该技术已经广泛应用于诸多领域，如公共安全、城市规划、商业、交通等基于位置的服务。

传统的空间并置模式挖掘方法将空间实例抽象成点对象，但规模较大的空间实例覆盖面积可能跨越很多实例的邻域，例如医院和花店、药店。因此，对于医院这样具有较大覆盖面积的实例，使用多个入口共同表示此实例更加准确。实际生活中，实例的位置因人们对其入口的认知而不同，具有模糊性。在生活实践中模糊数据普遍存在，数据的边界或概念的定义具有模糊性。比如，“高”和“矮”、“胖”和“瘦”、“远”和“近”等。现有的基于模糊集的空间并置模式挖掘方法用模糊隶属度来表示模糊数据[1]。本文使用隶属度集对多个入口隶属于同一模糊实例这一现象给出客观的度量方式，隶属度集为模糊实例不同入口处人流量大小的比例，以此来表示不同位置属于该实例的概率。

此外，传统的空间并置模式挖掘算法需要一个预定义的距离阈值来识别相互邻近的实例。此静态距离阈值方法操作简单，但如果要得到令用户满意的效果，需要用户根据主观经验和数据分布来反复实验寻找合适的阈值。因为不同特征空间分布密度存在差异，以及相同特征在不同区域也可能有不同的空间分布密度。“邻近”是一个相对的概念。例如，同样是位于 3000 米处的居民，对沃尔玛(空间分布密度小)而言是邻近，而对小区便利店(空间分布密度大)却很远。并且，现有算法忽略了邻近实例间的紧密程度，邻近实例在候选模式上的权重被平等看待，无法挖掘高精度的并置模式。比如距离阈值为 100 米时，100 米处的实例和 50 米处的实例对于模式的贡献是一样的，这显然是不合理的。

本文综合考虑了模糊实例规模、特征分布密度、邻近实例间的邻近度等因素对空间并置模式的影响，提出了基于 Voronoi 图和距离衰减效应的模糊实例空间并置模式挖掘方法。本文主要贡献如下：

1) 针对模糊实例规模对空间并置模式影响的问题，运用模糊理论中的隶属度度量同一实例不同入口对该实例位置的贡献，用各入口的隶属度计算实例规模，并将其融入实例间的邻近度计算中。

2) 针对邻近实例间的邻近度对并置模式影响的问题，首先，通过 Voronoi 图来识别邻近实例；然后，利用距离衰减函数对邻近实例间的邻近度进行更准确的计算。

3) 在真实数据集和合成数据集上实验验证本文的算法可以发现传统空间并置模式挖掘方法所未发现的有意义模式。

本文第二节给出相关工作，第三节介绍相关定义和性质，第四节分析算法流程和详细步骤，第五节展示实验结果和分析，第六节总结全文并讨论下一步工作方向。

2. 相关工作

空间并置模式挖掘算法最初由 Shekhar 和 Huang [2]提出，是一种使用 Apriori 策略的基于连接的算法。随后，许多研究人员开发了各种改进算法，并取得了令人满意的结果。其中一些算法，如部分连接算法[3]、无连接算法[4]、密度聚类算法[5]、基于顺序团的算法[6]、SGCT 算法[7]、杨培忠等提出的基于列计算的空间并置模式挖掘方法[8]和张绍雪等提出的避免逐阶挖掘的算法[9]，专注于提高效率。此后，学者们还提出了带稀有特征数据[10]、不确定数据[11]、核模式[12]和高效用[13]的空间并置模式挖掘算法，拓展了数据类型和挖掘目标。

由于空间实例位置及距离远近标定的模糊性，近年来，一些学者致力于模糊数据的空间并置模式挖掘的研究，提出了模糊对象的空间并置模式挖掘方法。文献[1]将模糊集与空间并置模式挖掘相结合，提出了模糊对象空间并置模式挖掘的相关定义和定理，并给出具体的算法和相应的剪枝策略。文献[14]将模糊实例位置的概念融入到空间并置模式挖掘中，并采用基于网格的距离计算方法来提高挖掘效率。文献[15]提出了一种挖掘模糊对象最大并置模式的 Mevent-Tree 算法。文献[16]提出了一种基于非均匀模糊空间对象的层次化并置模式挖掘方法，使得各层数据分布均匀，并将挖掘出的结果分成多层。

传统的空间并置模式挖掘过程中，空间实例被抽象成点对象，每个实例对应一个确定位置。然而，规模较大的空间实例有多个重要位置点(如医院、公园入口)，其空间位置因对其重要位置点的认知不同而存在差异，具有模糊性。对于这些模糊实例，本文考虑其重要位置点对该实例规模的贡献，将实例规模对空间并置模式挖掘中实例间的邻近程度的影响考虑了进来。

以上这些算法都基于静态的距离阈值来获取实例间邻近关系，过于依赖用户经验来寻找合适阈值。为了解决这一问题，一些学者提出了新的方法。例如，Wang 和 Zhou [17]提出了一种基于 k-最近特征的并置模式算法，其中实例之间的邻居关系取决于最近对象的数量和 k 值。然而，这些算法在实现之前仍然需要了解接近标准，例如 k 阈值。为了解决这一限制，已经开发了一些不需要邻近标准的自适应算法。例如，Sundaram 等人[18]使用 Delaunay 三角测量来寻找邻近实例。该方法将三角形边连接的节点视为邻

居。Qian 等人[19]提出了一个迭代框架来发现频繁的并置模式。该方法迭代选择有信息的边来构建邻居关系图，直到每个重要的模式都有足够的置信度。基于绝对和相对频繁度，Qian 等人[20]探索了一种采用 k 近邻图代替距离阈值发现区域并置模式的分层并置算法。

随着空间并置模式挖掘研究的不断深入，关于距离衰减效应对空间并置模式挖掘的影响得到关注。文献[21]指出实例对空间并置模式挖掘的影响会随着距离的增大而不断减少，并且提出一种考虑密度加权距离阈值的挖掘方法。文献[21]使用近邻的方法来识别邻近实例，考虑了实例的空间分布特点，但未对特征分布密度加以考虑。文献[22]通过 Voronoi 图对空间数据集的邻近关系进行识别，虽然考虑了特征分布密度，但对实例邻近程度的表达缺乏进一步的划分。本文考虑了实例位置的模糊性，结合 Voronoi 图与距离衰减函数，共同刻画邻近实例的邻近度，提出了基于 Voronoi 图和距离衰减效应的模糊空间并置模式挖掘算法。

3. 相关定义和性质

3.1. 相关工作定义和性质

本节介绍空间并置模式挖掘的相关定义和性质。**空间特征**是指空间中的事物(例如，医院，药店等)。空间特征集是空间中事物的集合，记为 $F=\{f_1,f_2,\cdots,f_n\}$。**空间实例**是空间特征在空间中具体位置的表示，记为 $S=\{I_1,\cdots,I_n\}$。I_t 是特征 f_t 的空间实例集合，$1\le t\le n$，实例表示为〈实例 ID，所属特征，实例位置〉。如果两个实例 o_i 和 o_j 之间的距离不大于用户给定的距离阈值 d，则称实例 o_i 和 o_j 满足**邻近关系** R，即 $R(o_i,o_j)\Leftrightarrow \rho(o_i,o_j)\le d$，$\rho$ 为实例间的欧式空间距离。$NR(o_i)=\{o_j|R(o_i,o_j)\wedge f(o_i)\ne f(o_j)\}$ 是满足邻近关系 R 的实例 o_i 的**邻居集**。$cl=(o_1,o_2,\cdots,o_k)$ 为空间中的一组实例，当 cl 中任意两个空间实例均满足空间邻近关系 R 时，称 cl 是一个**空间团**。当一个空间团不为任意空间团的子集时，称该空间团为一个**极大空间团**。

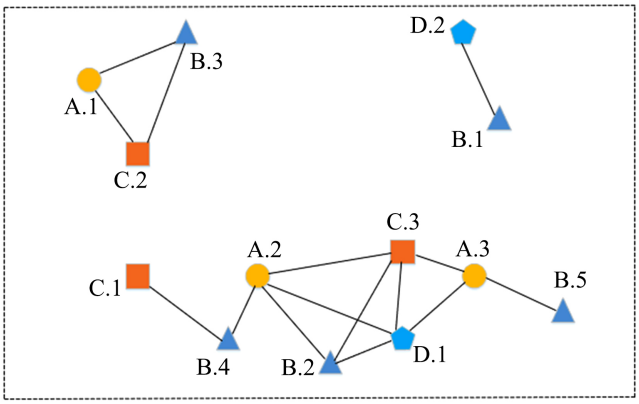


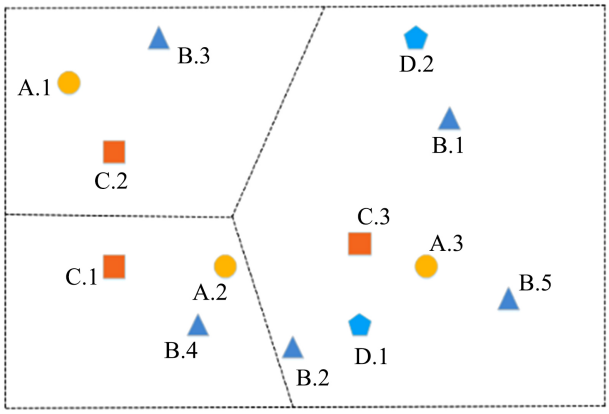
Figure 1. Example of a spatial data set
图 1. 空间数据集示例

空间并置模式 c 是空间特征集 F 的一个子集，$c=\{f_1,\cdots,f_k\}\subseteq F$。$c$ 中的特征数 k 称为 c 的阶。一个空间团 cl 包含了 c 中所有空间特征，且 cl 中任意两实例的特征互不相同，则称 cl 为 c 的一个**行实例**，c 的所有行实例构成了**表实例** $T(c)$。**参与率** $PR(c,f_i)$ 用来衡量特征 $f_i(1\le i\le k)$ 在模式 c 中的参与情况，$PR(c,f_i)=|f_i$在$T(c)$中的不重复实例的个数$|/|f_i$的实例总数$|$。空间并置模式 c 的所有空间特征中参与率的最小值为 c 的**参与度** $PI(c)$，即 $PI(c)=\min_{f_i\in c}\{PR(c,f_i)\}$。当 $PI(c)$ 不小于给定的频繁性阈值 $\min\_prev$ 时，称空间并置模式 c 是一个**频繁空间并置模式**。

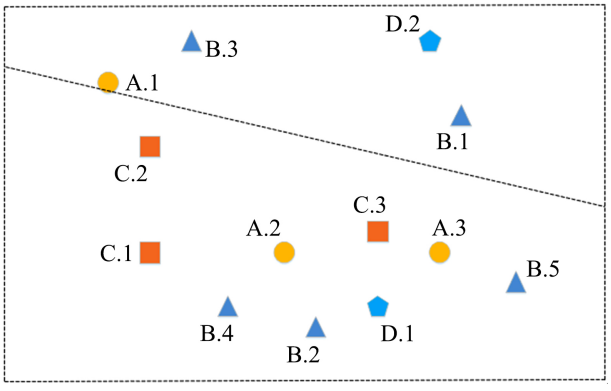
例 1. 图 1 为一个空间数据集示例，该数据集有 A、B、C、D 四个特征，其实例数分别为 3、5、3、2。图中将具有空间邻近关系的实例由实线连接。以模式{A, B, C}为例，其计算过程如下。表实例为 $\{\{A.1,B.3,C.2\},\{A.2,B.2,C.3\}\}$，$PR\{\{A,B,C\},A\}=|\{A.1,A.2\}|/|\{A.1,A.2,A.3\}|=2/3$，$PR\{\{A,B,C\},B\}=|\{B.2,B.3\}|/|\{B.1,B.2,B.3,B.4,B.5\}|=2/5$，$PR\{\{A,B,C\},C\}=|\{C.2,C.3\}|/|\{C.1,C.2,C.3\}|=2/3$，$PI(\{A,B,C\})=\min\{PR\{\{A,B,C\},A\},PR\{\{A,B,C\},B\},PR\{\{A,B,C\},C\}\}=0.4$。如果 min_ *prev* 小于或等于 0.4，则模式{A, B, C}是一个频繁空间并置模式。

为了更合理地度量空间并置模式挖掘中的距离衰减效应，考虑了特征实例分布密度，结合 Voronoi 图划分的邻近区域来定义一种距离衰减邻近度，细化了邻近实例间的邻近程度。Voronoi 图叫泰森多边形或维诺图，它是由一组由连接两邻点直线的垂直平分线组成的连续多边形。以空间数据集中某特征的各个实例为生成元构造泰森多边形，所得即为其他特征实例关于该特征实例的 Voronoi 图。在构造的 Voronoi 图各个分块区域中的其他特征实例都和该分块区域生成元的特征实例具有邻近关系。

空间实例集 I 关于特征 f 的 Voronoi **划分** $V(f)$ 是以特征 f 的实例集为生成元生成 Voronoi 图，由生成的 Voronoi 图对空间实例集 I 的划分称为空间实例集 I 关于特征 f 的 Voronoi 划分 $V(f)$。生成的 Voronoi 图中各块区域称为 Voronoi **块**。每一 Voronoi 块中其他特征的实例都与该 Voronoi 块生成元的特征实例具有邻近关系。Voronoi 块中关于该生成元特征 f 的**邻近实例对集为** $P(f)$。



(a) 以特征 C 的实例为生成元的 Voronoi 图



(b) 以特征 D 的实例为生成元的 Voronoi 图

Figure 2. Example of Voronoi partitioning of features
图 2. 特征的 Voronoi 划分示例

例 2. 图 2(a)是根据所给空间实例集 I 得到的关于空间特征 C 的 Voronoi 划分示例，以特征 C 的实例集为生成元，生成 Voronoi 图，得到 3 个 Voronoi 块，每个 Voronoi 块内的其他特征的实例和生成元实例具有邻近关系。Voronoi 块中关于生成元特征 C 的邻近实例对集 $P(\mathrm{C})=\{(\mathrm{C}.1,\mathrm{B}.4),(\mathrm{C}.2,\mathrm{A}.1),(\mathrm{C}.2,\mathrm{B}.3),\cdots,(\mathrm{C}.3,\mathrm{D}.2)\}$。图 2(b)是空间实例集 I 中关于空间特征 D 的 Voronoi 划分示例，以特征 D 的实例集为生成元得到 2 个 Voronoi 块，$P(\mathrm{D})=\{(\mathrm{D}.1,\mathrm{A}.2),(\mathrm{D}.1,\mathrm{A}.3),\cdots,(\mathrm{D}.2,\mathrm{B}.1),(\mathrm{D}.2,\mathrm{B}.3)\}$。

3.2. 本文工作定义和性质

定义 1. 实例间距离衰减邻近度设(o_i,o_j)为具有邻近关系的实例对，该实例对(o_i,o_j)的距离衰减邻近度定义如下：

$$D(o_i,o_j)=\mathrm{e}^{-\frac{dis(o_i,o_j)^2}{2(\max(d)^2)}} \tag{1}$$

$dis(o_i,o_j)$表示两实例之间的欧式距离，$\max(d)$表示 Voronoi 图划分邻近关系基础上邻近实例对的最大距离。$D(o_i,o_j)$的值域为(0, 1)，两实例越邻近，$D(o_i,o_j)$的值越接近于 1。

传统的空间并置模式挖掘过程中，将空间实例抽象成点对象，每个实例对应一个确定位置。然而，规模较大的空间实例有多个重要位置点(如医院、公园入口)，其空间位置因对其入口的认知不同而存在差异，具有模糊性。对于这些模糊实例，本文结合模糊理论使用隶属度集综合考虑隶属于该实例的多个重要位置点对该实例规模的贡献，重新定义实例间的邻近度。隶属度集为模糊实例不同重要位置点处人流量大小的比例，以此来表示不同位置隶属于该实例的概率。

定义 2. 模糊实例规模模糊实例表示为一个四元组〈实例 ID，所属特征，实例位置，隶属度集〉，其中隶属度集为模糊实例不同入口处人流量大小的比例。给定一个模糊实例o_m的隶属度集 $b(m)=(i_1,i_2,\cdots,i_t)$，其模糊实例规模定义如下：

$$S(o_m)=t\times\left(1-\frac{\min(b(m))}{\max(b(m))}\right)+0.01 \tag{2}$$

其中，$b(m)=(i_1,i_2,\cdots,i_t)$中的$i_j(1\le j\le t)$为第 j 处入口隶属于模糊实例的概率，其值为该入口人流量占该实例总人流量的比值，t 为模糊实例的入口数量。

在实际空间数据集中，入口数量 t 越多，实例规模往往越大。而具有多个入口的大规模实例，其各入口的人流量往往难以做到非常平均，即 $\min(b(m))$ 和 $\max(b(m))$ 相差较大，此时 $1-\frac{\min(b(m))}{\max b((m))}$ 较大。0.01 为避免当实例只有一个入口或 $\min(b(m))=\max(b(m))$ 时 $S(o_m)=0$。

定义 3. 模糊邻近实例对的效用度设(o_i,o_j)为具有邻近关系的模糊实例对，则该模糊实例对(o_i,o_j)的效用度为二者的模糊实例规模均值与距离衰减邻近度的乘积，定义如下：

$$SD(o_i,o_j)=avg\{S(o_i),S(o_j)\}\times D(o_i,o_j) \tag{3}$$

定义 4. 模式行实例中每个实例的效用度给定模式 c 的一个行实例 $cl(c)$，设 $o_i\in cl(c)$，为特征 f_i 的实例，则空间并置模式 c 中特征 f_i 的实例 o_i 的效用度为 $SD(c,o_i)$。

$$SD(c,o_i)=\sum_{\substack{o_i,o_j\in cl(c)\\ o_j.feture=f_j\wedge f_j\in c\setminus\{f_i\}}} SD(o_i,o_j) \tag{4}$$

定义 5. 实例规模距离衰减参与率空间特征 f_i 在空间并置模式 c 上实例规模距离衰减参与率 $SDPR(c, f_i)$ 表示为特征 f_i 在 c 的表实例上不重复出现实例的效用度的和与 f_i 所有模糊实例规模和的比。

$$SDPR(c, f_i) = \frac{\sum_{j=1}^{|f_j|}\left(SD(c, o_j)\right)}{S(f_i)} \tag{5}$$

$|f_j|$ 表示特征 f_i 在 c 的表实例上不重复出现实例个数，$S(f_i)$ 表示特征 f_i 所有实例规模的和，$S(f_i) = \sum_{i=1}^{|f_i|} S(o_i)$，$o_i \in f_i$，$|f_i|$ 表示特征 f_i 的总实例个数。

定义 6. 实例规模距离衰减参与度空间并置模式 $c = \{f_1, f_2, \cdots, f_n\}$ 上的实例规模距离衰减参与度 $SDPI(c)$ 表示为空间并置模式 c 上的最小的实例规模距离衰减参与率。

$$SDPI(c) = \min_{i=1}^{n}\{SDPR(c, f_i)\} \tag{6}$$

例 3. 对图 1 的空间数据集利用 Voronoi 图重新划分邻近关系得到实例间新的空间邻近关系如图 3 所示。以空间数据集中不同特征的各个实例为生成元构造多重泰森多边形，所得 Voronoi 图各个分块区域中的其他特征实例都和该分块区域生成元的特征实例具有邻近关系，利用 Voronoi 图对原数据集邻近关系进行了重新划分，将多重泰森多边形重新划分后的邻近关系进行取交集操作，即为最终结果。

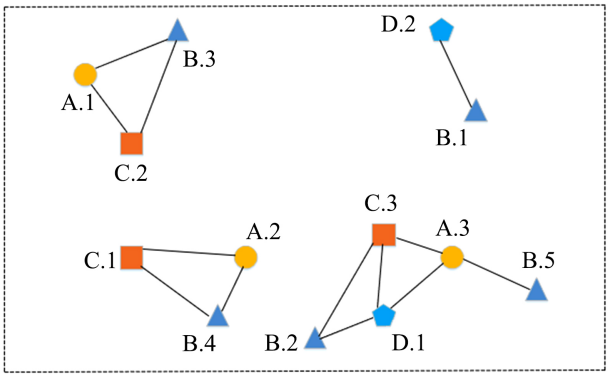


Figure 3. The proximity relationship of the spatial data set in Figure 1 was divided based on Voronoi diagram
图 3. 基于 Voronoi 图划分图 1 空间数据集邻近关系

表 1 为图 3 中各模糊实例的隶属度集，表 2 为根据图 1 中模糊实例的隶属度集利用定义 2 计算得到各模糊实例规模。表 3 为实例间距离及根据定义 1 得到实例间距离衰减邻近度和根据定义 3 得到邻近实例对的效用度的值。

Table 1. Membership degree set of fuzzy instance in Figure 3
表 1. 图 3 中模糊实例的隶属度集

实例	隶属度集	实例	隶属度集
A1	(0.3, 0.3, 0.4)	B4	(0.54, 0.46)
A2	(0.4, 0.6)	B5	(0.58, 0.42)
A3	(0.42, 0.58)	C1	(0.3, 0.4, 0.3)
B1	(0.46, 0.54)	C2	(0.6, 0.4)
B2	(0.38, 0.62)	C3	(0.62, 0.38)
B3	(0.4, 0.6)		

Table 2. Size of the fuzzy instance in Figure 3
表 2. 图 3 中模糊实例的规模

实例	A1	A2	A3	B1	B2	B3	B4	B5	C1	C2	C3
规模	0.76	0.35	0.57	0.31	0.79	0.35	0.31	0.57	0.76	0.35	0.79

Table 3. The distance between fuzzy instances, distance attenuation proximity and utility are shown in Figure 3
表 3. 图 3 中模糊实例间距离、距离衰减邻近度及效用度

	AB			AC			BC		
	A1B3	A2B4	A3B5	A1C2	A2C1	A3C3	B2C3	B3C2	B4C1
距离	3.8	3	3.1	3.6	4.3	3.2	5	5	4.7
$D\left(o_i,o_j\right)$	0.75	0.84	0.83	0.78	0.69	0.81	0.6	0.6	0.64
$SD\left(o_i,o_j\right)$	0.41	0.27	0.47	0.43	0.38	0.55	0.47	0.21	0.34

根据表 2、表 3 所给信息，利用定义 5 计算实例规模距离衰减参与率 $SDPR\left(c,f_i\right)$。候选并置模式 $c_1=\{\mathrm{B},\mathrm{C}\}$ 中，$SDPR\left(c_1,\mathrm{B}\right)=\left(0.47+0.21+0.34\right)/\left(0.31+0.79+0.35+0.31+0.57\right)=0.43$，$SDPR\left(c_1,\mathrm{C}\right)=\left(0.47+0.21+0.34\right)/\left(0.76+0.35+0.79\right)=0.53$，那么 $SDPI\left(c_1=\{\mathrm{B},\mathrm{C}\}\right)=\min\left\{SDPR\left(c_1,\mathrm{B}\right),SDPR\left(c_1,\mathrm{C}\right)\right\}=0.43$。同理，二阶模式 $c_2=\{\mathrm{A},\mathrm{B}\}$ 中的 $SDPI\left(c_2=\{\mathrm{A},\mathrm{B}\}\right)=\min\left\{SDPR\left(c_2,\mathrm{A}\right),SDPR\left(c_2,\mathrm{B}\right)\right\}=0.49$，对于三阶候选模式 $c_3=\{\mathrm{A},\mathrm{B},\mathrm{C}\}$，$SDPI\left(c_3=\{\mathrm{A},\mathrm{B},\mathrm{C}\}\right)=\min\left\{SDPR\left(c_3,\mathrm{A}\right),SDPR\left(c_3,\mathrm{B}\right),SDPR\left(c_3,\mathrm{C}\right)\right\}=0.62$。候选模式 $c_3=\{\mathrm{A},\mathrm{B},\mathrm{C}\}$ 的实例规模距离衰减参与度为 0.62，候选模式 $c_3=\{\mathrm{A},\mathrm{B},\mathrm{C}\}$ 的传统空间并置模式频繁度量方法使用模式在空间数据集中出现的频率作为参与度，计算其参与度为 0.4，说明本文所提算法能够挖掘到传统挖掘方法所忽略的将实例个体规模对于模式参与度的贡献考虑进来后的有意义模式。

引理 1. 实例规模距离衰减参与度不满足向下闭合性。

证明：在例 3 中，二阶模式{B, C}的实例规模距离衰减参与度 $SDPI\left(\{\mathrm{B},\mathrm{C}\}\right)$ 为 0.43，二阶模式{A, B}的实例规模距离衰减参与度为 0.49，三阶模式{A, B, C}的实例规模距离衰减参与度 $SDPI\left(\{\mathrm{A},\mathrm{B},\mathrm{C}\}\right)$ 为 0.62。若频繁阈值为 0.5，那么模式{B, C}、模式{A, B}不频繁，模式{A, B, C}频繁。低阶模式不频繁而高阶模式频繁，故不满足向下闭合性。

4. 基于极大团和哈希表的挖掘框架

基于引理 1，由于本文参与度度量方式不具有向下闭合性质，如果采用逐级搜索挖掘框架，不必要的候选模式不能被有效地剪枝，挖掘效率极低，特别是在数据集密集的情况下。此外，挖掘频繁并置模式的关键是收集模式的行实例，而这一步是最耗时的。因此，缩小候选搜索空间和快速收集行实例是提高挖掘算法效率的关键。本文使用极大团和哈希表的挖掘框架，基于空间实例的邻近关系搜索极大团，存储于双层哈希表中，快速获取模式的参与实例。避免了在生成－测试候选挖掘框架中对每个行实例的邻近关系进行验证。参与实例即出现在模式表实例中各特征的实例。

4.1. 搜索极大团步骤

从空间数据集中生成极大团是 NP 难问题，需要一种快捷的算法来生成空间数据集中的所有极大团，Bron-Kerbosch 算法[23]在用于枚举极大团方面效果良好。Eppstein 提供了 Bron-Kerbosch 算法的另一种改

进算法[24]，该算法对于低退化度的图获得了接近最优的最坏情况的时间耗费。先计算图的退化序列，然后 Eppstein 递归算法按退化序列顺序选择递归调用中的顶点 v，最后对于顺序中的每个顶点 v 依次计算极大团。所以本文选择该时间耗费更优的算法来获得极大团。

引理 2. 如果空间极大团 *cl* 是空间并置模式 *c* 的一条行实例，则从 *cl* 中可以得到模式 *c* 及其所有子集的行实例。

证明：假设有空间并置模式 *c* 和子集 c'，$c' \subseteq c$。若极大团 *cl* 是 *c* 的一条行实例，因 *cl* 中任意两实例均邻近，则 *cl* 中包含 c' 特征的子团 cl' 是 c' 的一条行实例。

由引理 2 可知，一个模式的行实例，可从它及其超集对应的极大团中获得。模式表实例的计算不再像传统空间并置模式挖掘算法那样，依赖低阶模式的表实例逐阶生成候选并测试。

4.2. 双层哈希表挖掘空间并置模式步骤

本文使用双层哈希表来储存极大团，其形式为$\left(key_1,\left(key_2,value_2\right)\right)$，$key_1$ 是极大团中实例的空间特征的集合，key_2 是极大团中每个实例的空间特征，$value_2$ 是极大团中实例自身。所有极大团中属于相同特征集的实例被分到双层哈希表中的同一个 $value_2$ 中。给定一个极大团，先检查哈希表中某个 key_1 与极大团特征组是否相同，若相同则将极大团的特征和实例放入该哈希表的$\left(key_2,value_2\right)$中。若不相同，则新建一个节点，将极大团的特征组作为该节点的 key_1，极大团的各特征及实例作为该节点的$\left(key_2,value_2\right)$。

例 4. 图 4 是基于图 1 数据集中极大团构建是双层哈希表 *CIHash* 的示例，空间极大团$\{A.2,B.2,C.3,D.1\}$的键 key_1 为{A, B, C, D}，值 $value_1$ 为$\{\langle A,A.2\rangle,\langle B,B.2\rangle,\langle C,C.3\rangle,\langle D,D.1\rangle\}$，值 $value_1$ 中的键 key_2 分别为{A}、{B}、{C}、{D}，值 $value_2$ 分别为{A.2}、{B.2}、{C.3}、{D.1}。

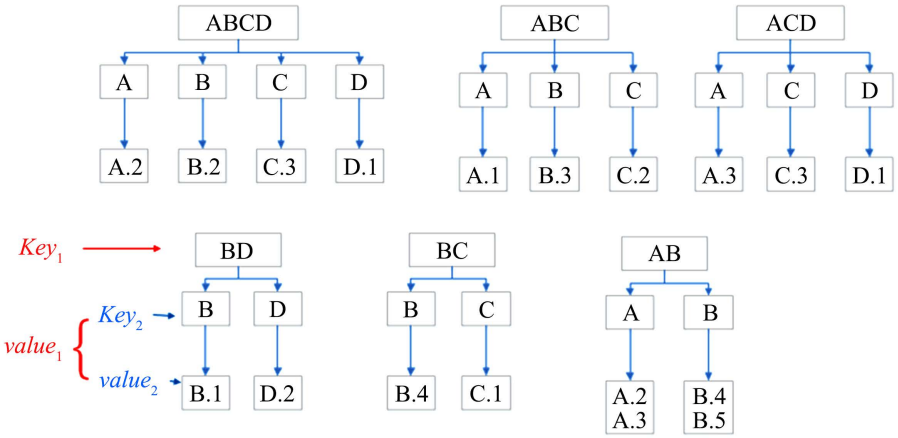


Figure 4. Build a two-layer hash table *CIHash* example based on the maximal clique of the data set in Figure 1
图 4. 基于图 1 数据集中极大团构建双层哈希表 *CIHash* 示例

双层哈希表中存储了极大团，如何从哈希表结构中快速查找参与每个模式的参与实例呢？首先，由双层哈希表中的键得到候选模式(即团候选)，然后从并置实例哈希表的值获得候选模式中特征的参与实例，计算参与度。

引理 3. 给定一个并置模式 *c*，从 *c* 及其超集的键 key_1 对应的双层哈希表的值 $value_2$ 中可以搜集到参与 *c* 的表实例中每个特征的参与实例。

证明：根据双层哈希表的定义，$value_2$ 是在模式 key_1 中属于特征 key_2 的一组参与实例集。如果模式 *c*

恒等于 key_1，那么可以从 $value_2$ 直接获得模式 c 的参与实例。如果模式 c 的参与实例是用双层哈希表储存极大团的子集，则模式 c 的参与实例可以通过查找模式 c 的超集的键对应的值来获得。

4.3. 算法描述

算法 1：基于 Voronoi 图和距离衰减效应的模糊实例空间并置模式挖掘算法

输入：Voronoi 图划分邻近关系后的空间数据集 $G(V, E)$
输出：频繁并置模式集 *CPS*
变量：双层哈希表 *CIHash*、频繁阈值 u、所有极大团 *MCs*

1. *MCs=Bron-Kerbosch-Degeneracy*(*G*(*V*,*E*))
2. *CIHash*←*MCs*
3. *keyset*←*CIHash.getKeys*()
4. *whilekeyset*≠∅*do*
5. *keyset*←*sort_by_size*(*keyset*)
6. *c*←*keyset.pop*()
7. *foritem* ∈ *CIHashdo*
8. *ifc* ⊆ *item.getKey*()*then*
9. *T*(*c*).*add*(*item.getValue*)
10. *endif*
11. *endfor*
12. *SDPI*(*c*)←*calculate_SDPI*(*c*,*T*(*c*))
13. *ifSDPI*(*c*) ≥ *uthen*
14. *CPS.add*(*c*)
15. endif
16. *sub_c*←*generate_subsets*(*c*)
17. *keyset*←*keyset* ⋃ *sub_c*
18. *endwhile*
19. *returnCPS*

算法过程：首先调用基于退化的 Bron-Kerbosch 算法来获取极大团，再将计算得到的极大团放进双层哈希表中存储(Step 1~2)，然后获取双层哈希表中的所有键即候选模式，并将它们存储在集合 *keyset* 中(Step 3)。算法执行 *while* 循环对候选模式进行检验。在循环中，将 *keyset* 按模式大小降序排序(Step 4~5)，取出来第一个键作为并置模式 c (Step 6)。查询键为 c 及其超集的哈希表节点(Step 8)，获取其值放入 c 的表实例中(Step 9)。然后计算 c 的实例规模距离衰减效用度(Step 12)。如果 c 参与度大于给定频繁阈值，将其放在频繁并置模式结果集上(Step 13~14)，生成 c 的子集并进 *keyset* 集(Step 16~17)。最后，将一组频繁并置模式集返回给用户(Step 19)。

4.4. 算法分析

4.4.1. 时间复杂度

算法时间复杂度主要包含极大团物化，候选模式过滤等部分。有学者研究[24]表明，顶点个数为 n 的无向图 G 中，带轴 *Bron-Kerbosch* 算法的最坏运行时间为 $O\left(3^{n/3}\right)$，所以基于退化度的极大团挖掘的时间复杂度为 $O\left(km3^{k/3}\right)$。k 为二阶频繁并置模式的退化度，m 是特征的数量。用双层哈希表储存极大团的

时间复杂度为$O(|MC|)$ ($|MC|$为极大团集合)。候选模式过滤的时间复杂度为$O(l \times m_{avg} \times 2^m)$，$l$是候选模式的数量，$m_{avg}$是候选模式的平均长度。

4.4.2. 空间复杂度

本文提出的算法主要空间耗费为哈希结构 *CIHash* 和候选模式参与实例的存储。设w_{avg}为极大团的平均长度，极大团数量为$|MC|$，那么哈希结构的存储耗费约为$O(w_{avg} \times |MC|)$。所有团候选参与实例的存储耗费约为$O(v_{avg} \times w_{avg} \times |MC|)$，$v_{avg}$是团候选中参与实例的平均长度。所以本文提出算法的空间耗费约为$O(v_{avg} \times w_{avg} \times |MC|)$。

4.4.3. 完备性和正确性

完备性：因为空间实例的邻近关系被极大团完整保存，双层哈希表不会遗漏邻近关系。算法 1 计算了所有模式的参与度，引理 2，引理 3 确保了算法 1 完备性，所以本文算法能挖掘到所有频繁空间并置模式。

正确性：由引理 3 可知，从双层哈希表中可以正确计算模式参与度，算法 1 可以保证计算的空间并置模式只有参与度满足频繁阈值才被放进结果，所以本文算法挖掘的均为满足频繁阈值的模式。

5. 实验

为了验证本文算法的实际效果和运行效率，将本文提出的算法(算法 1)和文献[4]提出的传统算法 Joinless 算法(算法 2)在挖掘到的频繁模式数量和运行时间上进行了对比实验。通过实验结果可以得到，我们的算法相比较传统挖掘算法能够挖掘到更多数量的频繁模式且运行时间更短，可以发现传统空间并置模式挖掘方法所忽略的将实例个体规模对于模式参与度的贡献考虑进来后的有意义模式。

本实验算法均使用 Python 编写，在电脑配置为 Win10 系统、4GB 内存的实验环境上运行获得结果。

5.1. 真实数据集

本实验所使用的真实数据为云南某地区的植被分布数据，该数据集包含 9 种植物种类，实例数量共有 28,783 个。在基于频繁阈值和距离阈值变化的情形下，进行对比实验比较两种算法在挖掘到的频繁模式数量和运行时间的结果。图中，蓝色实线表示算法 1 的运行结果，橙色虚线表示算法 2 的运行结果。

5.1.1. 频繁阈值 min_prev 的变化

本节实验参数为在距离阈值 d 为 15 m 的情形下，最小频繁阈值在 0.45 到 0.65 之间以步长为 0.05 进行变化的对比实验，比较二者挖掘到的频繁模式数量及运行时间。

图 5 展示了在距离阈值不变，最小频繁阈值不断增大的情形下，算法 1 和算法 2 所能挖掘到的频繁模式数量不断下降，算法 1 挖掘到的频繁模式数量明显多于算法 2，其中在频繁阈值为 0.5 时差距最为明显。图 6 展示了在距离阈值不变，最小频繁阈值不断增大的情形下，两算法进行频繁模式挖掘所需要的时间不断减少。频繁阈值影响模式数量的多少，模式的行实例增多，算法 2 查找表实例耗费时间长，算法 1 查找参与实例耗费时间短。从图中也可以看出算法 1 的运行时间明显小于算法 2，其中频繁阈值为 0.45 时差距最为明显。

5.1.2. 距离阈值 d 的变化

本节实验参数为在最小频繁阈值为 0.6 的情形下，距离阈值在 14 m 到 18 m 之间以步长为 1 m 进行变化对比实验，比较二者挖掘到的频繁模式数量及运行时间。

图 7 展示了不同距离阈值下两种算法挖掘到的频繁模式数量的折线图。从图中可以得到在距离阈值

为 14 m 处挖掘到的频繁模式数量相近，随着距离阈值增大，两种算法挖掘到的模式数量不断增多，模式数量差异逐渐增大。图 8 展示了不同距离阈值下两种算法挖掘算法运行时间的折线图。在距离阈值为 14 m 处二者运行时间接近，随着距离阈值的增大，邻近实例和极大团的数量增多，两算法的运行时间耗费增大，在距离阈值 16 m 处之后差距明显。

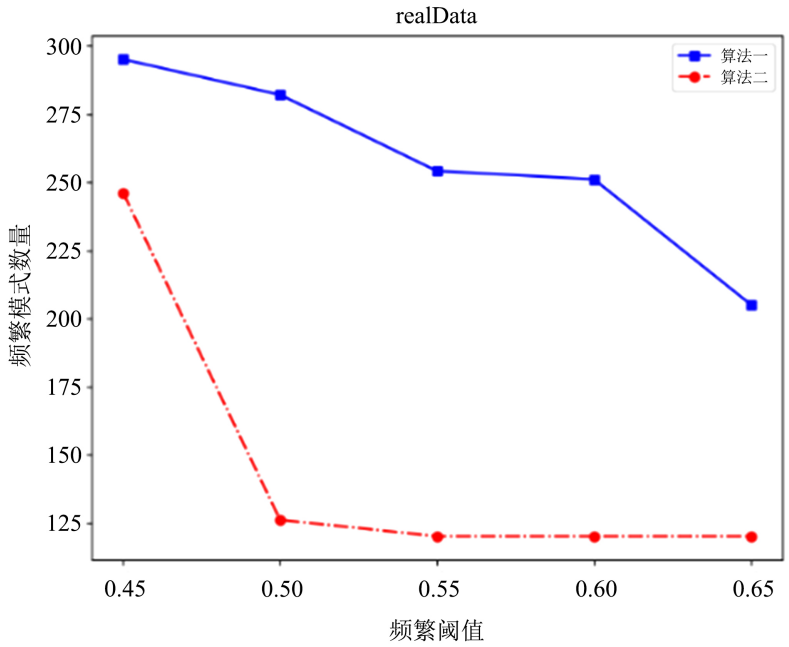


Figure 5. The effect of frequency threshold on the number of patterns in real data
图 5. 真实数据中频繁阈值对模式数量的影响

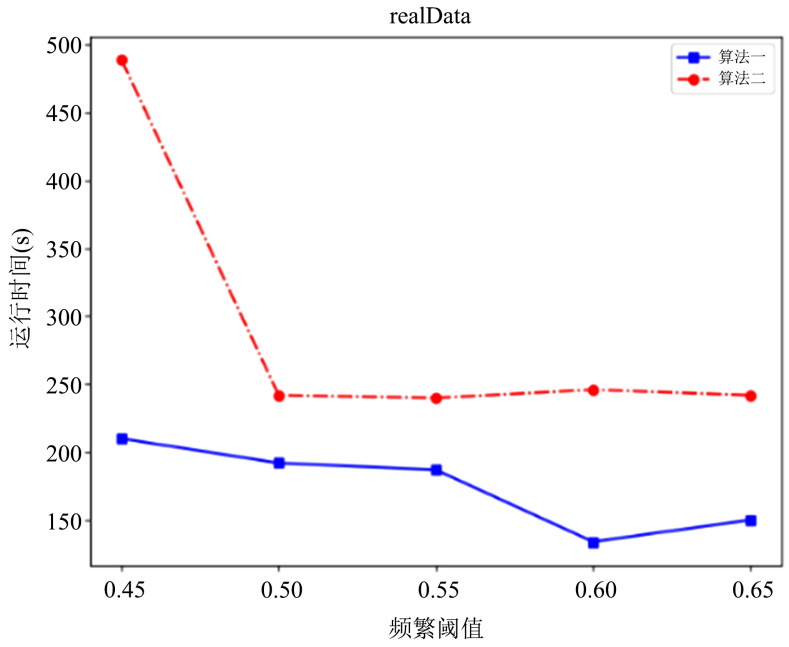


Figure 6. The effect of frequency threshold on run time in real data
图 6. 真实数据中频繁阈对运行时间的影响

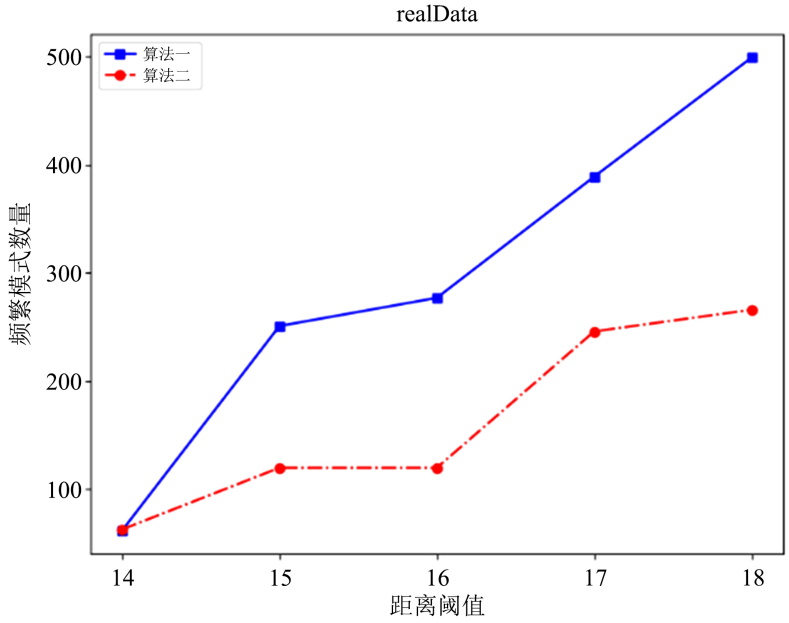


Figure 7. The influence of distance threshold on the number of modes in real data
图 7. 真实数据中距离阈值对模式数量的影响

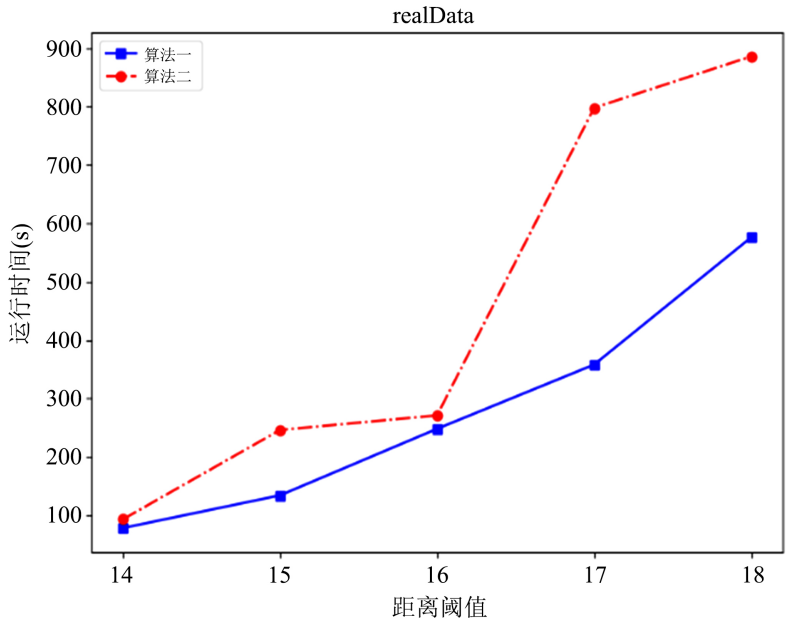


Figure 8. The influence of distance threshold on running time in real data
图 8. 真实数据中距离阈值对运行时间的影响

5.2. 合成数据集

本节实验使用合成数据的特征数量为 10，实例数量总数设定为 59,994 个，各个特征下的实例数量随机生成。实验对比结果和变化参数如下。

频繁阈值和距离阈值的变化

图 9 展示了两个算法在频繁阈值逐渐增大的情形下所能挖掘到的频繁模式数量的变化趋势。从图中

可以得到，在距离阈值为 0.6 到 0.7 时二者的变化趋势都较为平稳，二者的挖掘结果差距明显。随着频繁阈值的增大，挖掘到的模式数量均减少。图 10 展示了两个算法在距离阈值逐渐增大的情形下所能挖掘到的频繁模式数量的变化趋势。从图中可以得到，在距离阈值为 13、14 m 处时，两个算法挖掘的模式数量差距相对稳定，而随着距离阈值的增大，算法 1 挖掘到的模式数量增速明显，而算法 2 增速缓慢。

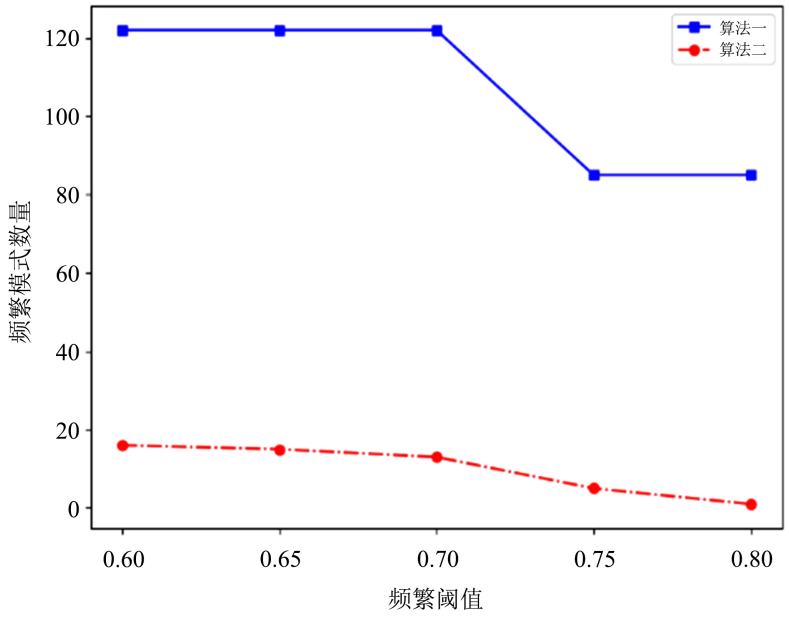


Figure 9. The effect of frequency threshold on the number of patterns in synthetic data
图 9. 合成数据中频繁阈值对模式数量的影响

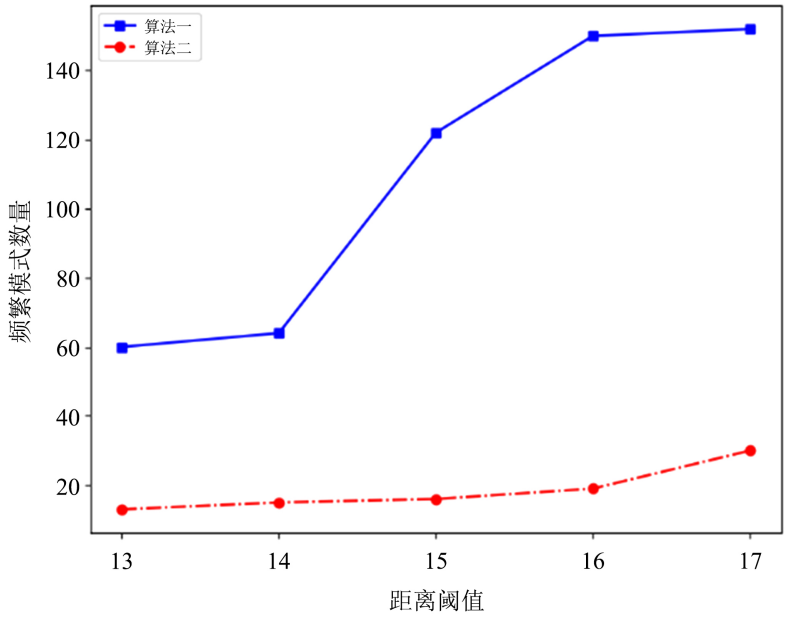


Figure 10. The influence of distance threshold on the number of patterns in synthetic data
图 10. 合成数据中距离阈值对模式数量的影响

通过本节的实验，可以得到本文的算法相比较传统挖掘算法能够挖掘到传统空间并置模式挖掘方法所忽略的将实例个体规模对于模式参与度的贡献考虑进来后的有意义模式且运行时间更短，在实验结果的折线图上有了清晰的展示。

6. 结束语

本文考虑空间实例的重要位置点(例如公园、医院不同入口)对该实例规模的贡献，重新定义实例间的邻近度。此外，考虑了空间特征分布密度，结合维诺图和距离衰减函数，自适应地确定邻近关系并刻画邻近实例间的邻近度。提出基于维诺图和距离衰减效应的模糊实例空间并置模式挖掘方法。为实现快速挖掘，设计了极大团和哈希表搜索参与实例的挖掘框架。在真实数据集和合成数据集上进行实验，验证本文的算法可以发现传统空间并置模式挖掘方法所忽略的有意义模式。

参考文献

[1] Ouyang, Z., Wang, L. and Chen, H. (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

[2] Shekhar, S. and Huang, Y. (2001) Discovering Spatial Co-Location Patterns: A Summary of Results. In: *International Symposium on Spatial and Temporal Databases*, Springer, Berlin, 236-256. https://doi.org/10.1007/3-540-47724-1_13

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

[4] Yoo, J.S. and Shekhar, S. (2006) A Joinless Approach for Mining Spatial Colocation Patterns. *IEEE Transactions on Knowledge and Data Engineering*, **18**, 1323-1337. https://doi.org/10.1109/TKDE.2006.150

[5] Huang, Y. and Zhang, P. (2006) On the Relationships between Clustering and Spatial Co-Location Pattern Mining. 2006 18*th IEEE International Conference on Tools with Artificial Intelligence* (*ICTAI*'06), Arlington, 13-15 November 2006, 513-522. https://doi.org/10.1109/ICTAI.2006.91

[6] Wang, L., Zhou, L., Lu, J., *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

[7] Yao, X., Peng, L., Yang, L., *et al*. (2016) A Fast Space-Saving Algorithm for Maximal Co-Location Pattern Mining. *Expert Systems with Applications*, **63**, 310-323. https://doi.org/10.1016/j.eswa.2016.07.007

[8] 杨培忠, 王丽珍, 王晓璇, 等. 一种基于列计算的空间并置模式挖掘方法[J]. 中国科学(信息科学), 2022, 52(6): 1053-1068.

[9] 张绍雪, 王丽珍, 陈文和. CPM-MCHM: 一种基于极大团和哈希表的空间并置模式挖掘算法[J]. 计算机学报, 2022, 45(3): 526-541.

[10] 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

[11] Wang, L., Wu, P. and Chen, H. (2011) Finding Probabilistic Prevalent Colocations in Spatially Uncertain Data Sets. *IEEE Transactions on Knowledge and Data Engineering*, **25**, 790-804. https://doi.org/10.1109/TKDE.2011.256

[12] 罗金, 王丽珍, 王晓璇, 等. K-近邻关系下的空间高效用核模式挖掘[J]. 计算机学报, 2022, 45(2): 354-368.

[13] Wang, L., Jiang, W., Chen, H., *et al*. (2017) Efficiently Mining High Utility Co-Location Patterns from Spatial Data Sets with Instance-Specific Utilities. *Database Systems for Advanced Applications*: 22*nd International Conference, DASFAA* 2017, Suzhou, 27-30 March 2017, 458-474. https://doi.org/10.1007/978-3-319-55699-4_28

[14] Ouyang, Z., Wang, L. and Zhou, L. (2012) Mining Spatial Co-Location Patterns for Fuzzy Location of Instances. *Journal of Frontiers of Computer Science & Technology*, **6**, 1144-1152.

[15] Wen, F., Xiao, Q. and Wang, L. (2014) Algorithm of Mining Maximal Co-Location Patterns for Fuzzy Objects. *Computer Science*, **41**, 138-145.

[16] Yu, Q., Luo, Y., Wu, Q., *et al*. (2016) Hierarchical Co-Location Pattern Mining Approach of Unevenly Distributed Fuzzy Spatial Objects. *Journal of Computer Applications*, **36**, 3113-3117.

[17] Wan, Y. and Zhou, J. (2008) KNFCOM-T: A K-Nearest Features-Based Co-Location Pattern Mining Algorithm for Large Spatial Data Sets By Using T-Trees. *International Journal of Business Intelligence and Data Mining*, **3**, 375-389. https://doi.org/10.1504/IJBIDM.2008.022735

[18] Sundaram, V.M. and Paneer, P. (2012) Discovering Co-Location Patterns from Spatial Domain Using a Delaunay Approach. *Procedia Engineering*, **38**, 2832-2845. https://doi.org/10.1016/j.proeng.2012.06.332

[19] Qian, F., He, Q., Chiew, K., *et al*. (2012) Spatial Co-Location Pattern Discovery without Thresholds. *Knowledge and Information Systems*, **33**, 419-445. https://doi.org/10.1007/s10115-012-0506-9

[20] Qian, F., Chiew, K., He, Q., *et al*. (2014) Mining Regional Co-Location Patterns with K NNG. *Journal of Intelligent Information Systems*, **42**, 485-505. https://doi.org/10.1007/s10844-013-0280-5

[21] Yao, X., Chen, L., Peng, L., *et al*. (2017) A Co-Location Pattern-Mining Algorithm with a Density-Weighted Distance Thresholding Consideration. *Information Sciences*, **396**, 144-161. https://doi.org/10.1016/j.ins.2017.02.040

[22] Yao, X., Chen, L., Wen, C., *et al*. (2018) A Spatial Co-Location Mining Algorithm That Includes Adaptive Proximity Improvements and Distant Instance References. *International Journal of Geographical Information Science*, **32**, 980-1005. https://doi.org/10.1080/13658816.2018.1431839

[23] Bron, C. and Kerbosch, J. (1973) Algorithm 457: Finding All Cliques of an Undirected Graph. *Communications of the ACM*, **16**, 575-577. https://doi.org/10.1145/362342.362367

[24] Eppstein, D., Löffler, M. and Strash, D. (2013) Listing All Maximal Cliques in Large Sparse Real-World Graphs. *Journal of Experimental Algorithmics* (*JEA*), **18**, 3.1-3.21. https://doi.org/10.1145/2543629