k-近邻关系下的空间高效用核模式挖掘

罗 金 王丽珍 王晓璇 肖 清
(云南大学信息学院 昆明 650504)

摘 要 空间数据挖掘旨在从空间数据库中发现和提取有价值的潜在知识.空间 co-location(共存)模式挖掘一直以来都是空间数据挖掘领域的重要研究方向之一,其目的是发现一组频繁邻近出现的空间特征子集,而空间高效用 co-location 模式挖掘则考虑了特征的效用属性.二者在度量空间实例的邻近关系时一般都需要预先给定一个距离阈值 d,挖掘算法的效率很受距离阈值 d 的影响,尤其对分布不均匀的数据集表现不好.另外,传统的空间高效用模式挖掘在分析评估模式的效用时,将模式中所有特征的效用值都计算到模式效用中是不合理的,如在国内 5A 级景区周围进行高收益商业项目的规划时,项目的预期收益本身不应包含景点的收益.基于上述问题,本文在空间高效用 co-location 模式挖掘过程中融入了空间 k-近邻计算,使得空间实例之间的邻近关系更为客观、合理.进一步地,定义了核元素和核模式等概念,对核模式效用的高低进行了度量,并提出了 k-近邻关系下的空间高效用核模式挖掘的通用框架,设计了一个行之有效的基本挖掘算法,考虑到核模式效用度不满足反单调性质,在基本算法之上提出了 4 个剪枝策略.大量的实验结果表明本文方法挖掘到的空间高效用核模式更具有现实意义,在同等的参数设置下,剪枝优化算法的效率比基本算法至少提高了 50%.

关键词 空间数据挖掘;空间 co-location 模式;空间高效用核模式;k-近邻
中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2022.00354

Mining Spatial High Utility Core Patterns under k-Nearest Neighbors

LUO Jin WANG Li-Zhen WANG Xiao-Xuan XIAO Qing
(*School of Information Science and Engineering, Yunnan University, Kunming* 650504)

Abstract Spatial data mining aims to help people discover and extract valuable patterns and knowledge from spatial data sets. Spatial co-location pattern mining has always been one of the important research directions in the field of spatial data mining, intending to find subsets of spatial features that often appear close together, while spatial high utility co-location pattern mining takes into account the utility attributes of the features. When measuring the neighbor relationship between spatial instances, both of the above two mining methods usually require a user setting distance threshold of d, the efficiency and effect of the mining algorithms are greatly affected by the distance threshold d. In particular, such algorithms do not work well on unevenly distributed datasets. In addition, when analyzing the pattern utility in the traditional spatial high utility co-location pattern mining, users are not interested in the utility value of some features in a pattern, which should not be calculated into the pattern utility together. Such as when planning of commercial project around 5A-level scenic spots in China to obtain high-yield returns, the expected income of the project itself should not include the income of the scenic spots. That is the spatial high utility pattern obtained by the traditional spatial high utility co-location pattern mining method is not necessarily reliable. Based on the above problems, this paper introduces the

收稿日期:2020-08-31;在线发布日期:2021-04-13.本课题得到国家自然科学基金项目(61966036,61662086)、云南省创新团队项目(2018HC019)资助.**罗 金**,硕士研究生,主要研究方向为空间数据挖掘.E-mail:luojin9420@163.com.**王丽珍**(通信作者),博士,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为空间数据挖掘、交互式数据挖掘、大数据分析等.E-mail:lzhwang@ynu.edu.cn.**王晓璇**,博士研究生,主要研究方向为空间数据库、数据挖掘.**肖 清**,硕士,讲师,主要研究方向为数据挖掘及应用.

k-nearest neighbor relationship into the spatial high utility co-location pattern mining. While solving the problem of setting the distance threshold, making the neighbor relationship between spatial instances is more objective and reasonable. Further, the concepts of core elements and core patterns are defined in the paper. In order to measure the utility of the proposed core pattern, the core participation instance set of features in a core pattern, the core utility participation ratio of features in a core pattern and the core utility index of a core pattern are formally defined. The problem that some feature utilities should not be included in spatial high utility co-location pattern mining is solved efficiently. Then a general mining framework for spatial high utility core pattern mining under the k-nearest neighbor relationship is proposed, and a basic mining algorithm is designed. In the basic algorithm, a grid-based method is used to replace the traditional brute force method for calculating the distance between spatial instances to obtain the k-nearest neighbor instance set of the instance, and the method named sequence tree is used to replace the fully joined method in the traditional spatial co-location pattern mining to quickly collect candidate patterns. In addition, considering that the utility of the core pattern does not satisfy the downward closure property, this paper presents four pruning strategies to greatly improve the mining efficiency of the basic algorithm. The time complexity, space complexity, completeness and correctness of the algorithm are analyzed. Finally, extensive experimental results on real and synthetic data sets show that the spatial high utility core patterns mined by this paper method are useful, with the pruning optimization algorithm being at least 50% better than the basic algorithm under the same parameter setting.

Keywords　spatial data mining; spatial co-location pattern; spatial high utility core pattern; k-nearest neighbor

1　引　言

在全球卫星定位系统和遥感技术等迅速发展的今天，如何从海量的空间数据中发现和提取有“价值”的知识已经是各个国家空间数据挖掘人员的主要任务之一. 空间共存(co-location)模式挖掘是空间数据挖掘的一个重要分支，旨在从空间数据中提取那些频繁邻近出现的空间特征的子集，在城市规划[1]、生态系统[2-3]和交通领域[4-5]等都有广泛的应用. 但研究人员逐渐发现，空间数据除了具有空间属性外，还具有一些不可忽视的非空间属性，如价值、时间等；价值属性也称为效用属性，是事物或数据对于不同的使用者或领域其重要性程度度量的指标，如商品的价格对于商业经济来说就是一个效用值，学校的教学评估等级(如 A、B、C、D 等)就是其在教学质量方面的效用值等；“效用”一词在数据挖掘领域最初出现在基于事务数据的高效用项集挖掘中[6]，其以香水带动口红和钻石销售的例子说明具有较高应用价值的项不一定是频繁出现的. 空间高效用 co-location 模式挖掘就是引入了空间数据的效用属性到空间模式的兴趣性度量中，结果发现空间数据带效用后，原本频繁的空间模式不一定具有较高的效用，而原来被忽视的一些模式却成了高效用的空间 co-location 模式，从而提出了各种空间高效用模式挖掘技术[7-10]，弥补了 co-location 模式挖掘中知识的遗漏和其隐含的预测性信息.

空间数据本质上都具有效用属性，只是效用的定义和高低不同，区别于经典事务数据库中对于效用的理解和定义，空间数据的效用很难用统一的标准来衡量，对应不同的领域和应用需求应有不同的定义，例如在生态系统中珍稀动植物的保护级别可以作为效用属性，在卫生健康领域中草药的药用价值可以作为效用属性，在商业活动中钻石、黄金等的经济价值是效用属性等. 所以基于事务数据的高效用模式挖掘方法并不适用于空间高效用模式挖掘.

现有的空间高效用 co-location 模式挖掘方法可分为两大类：考虑空间特征带效用的挖掘方法和考虑空间实例带效用的挖掘方法. 由于现实中空间实例分布的差异性，实例带效用的挖掘框架更符合

实际意义.如商业活动中空间特征为钻石,由于在自然界中的钻石大小不一和成色差异等,它的实例效用是不一样的.在城市规划中特征为酒店、学校、小区、高铁站等,每一个特征的实例各有差异,生态系统中树木的大小,药材的年份等不同也使得同一特征的不同实例效用不同.

空间高效用模式挖掘是从经典的空间模式挖掘演化发展而来,所以在度量空间实例的邻近关系的时候,沿用了传统的方法,通常是人为指定单一的距离阈值 d,计算出实例之间的距离,小于或等于 d 的实例对被认为是邻近的,这种度量方式存在以下几个问题:

(1)人为指定单一的距离阈值缺乏科学性.空间实例的分布大多是随机自然产生的,如两种植物之间的距离、河流与山脉的距离等.

(2)算法对 d 比较敏感.可能给定了一个较大的阈值 d,算法性能降低的同时挖掘到了大量模式;如果减小阈值 d,有可能算法性能提高了,但挖掘不到有意义的模式.

(3)挖掘框架不适用于不均匀的数据集.由于空间数据分布的差异性,当指定 d 时,会在分布稠密的区域得到较多的候选模式,但是在数据分布稀疏的区域得到的候选模式就比较少,所以也很难挖掘出带稀有特征的模式.

传统的空间高效用 co-location 模式挖掘基本都采用了一种类 PI(参与度)的挖掘框架,即求出每个特征在候选模式中的效用参与率,取最小的一个跟效用度阈值比较,大于等于效用度阈值的候选模式即是高效用模式.然而值得注意的是,在一些实际应用中,求出每个特征的效用参与率并不是得到空间高效用模式的必要条件.如在国内的 5A 级景区周围进行商业规划,挖掘那些和景区一起出现的高效用模式,采用传统的高效用模式挖掘方法,会将景区本身的效用值(如旅游收益)一起计算到模式效用中,5A 景区的收益一般都是比较高的,这样便会得到许多不可靠的"高效用模式".就商业规划而言,投资者更关心的是和景区邻近的拟投资的商业体(如酒店,餐饮行业)组成的模式是否是高效用的模式.

基于以上问题,本文提出了一种"k-近邻关系下的空间高效用核模式挖掘"的挖掘框架.主要贡献如下:

(1)将 k-近邻引入到空间高效用模式挖掘中,更加合理地度量了空间实例之间的邻近关系,对于均匀和不均匀分布的空间数据集,挖掘方法都具有良好的可伸缩性.

(2)提出了 k-近邻关系下核元素和核模式的概念,解决了传统的空间高效用 co-location 模式挖掘中将一些用户不感兴趣的特征的效用值一起计算到模式中的问题,同时设计了空间高效用核模式的挖掘算法并进行了优化.

(3)在真实数据集和模拟数据集上进行了广泛的实验和深入的分析,验证了所提出算法的正确性和有效性,优化算法的优化效果明显,挖掘结果也更符合实际应用,并且能挖出带有稀有特征的模式.

本文第 2 节介绍相关工作;第 3 节在给出空间高效用核模式的相关定义后,分析其具有的一些性质;第 4 节详细介绍 k-近邻关系下空间高效用核模式挖掘的基础算法和优化算法;第 5 节给出实验结果和分析;第 6 节总结全文并讨论未来工作.

2 相关工作

空间数据挖掘旨在从空间数据库中建模分析来帮助用户或科研做决策,历经多年的发展取得了显著的成果,其中比较有代表性的著作如文献[11]结合了计算机与地理信息科学,提供了系统和实用的空间数据挖掘概况.空间 co-location 模式挖掘作为空间数据挖掘领域的重要研究方向之一,其概念与定义最早于文献[12]中被提出.除了经典的 co-location 模式挖掘算法外,近年来国内外学者又做了大量 co-location 模式挖掘相关方面的研究.例如文献[13]针对现有大数据平台对空间数据挖掘缺乏全面支持的问题,提出一种面向大数据的空间模式挖掘框架.文献[14]将时间维度考虑到模式挖掘中,以空气污染的影响分析为案例对区域时空模式挖掘进行了研究,提出了一种通用的区域时空共存模式挖掘框架.文献[15]提出了一种新颖的基于团的方法获取完整和正确的 co-location 频繁模式,避免了需要验证行实例的问题,使得算法更容易找到一个合适的距离阈值.其他方面的研究,如文献[16]基于星型参与实例提出了"亚频繁 co-location 模式"的新概念,并设计了两个有效的算法挖掘极大亚频繁模式;文献[17]考虑到传统 co-location 模式挖掘中模式实例重叠性的问题,提出了一种"fraction-score"的新度量方法等.以上提到或未提到的大多数空间模式挖掘方法都是通过设定单一的距离阈值来进行空间邻近关系的判断,并且基于参与度阈值来衡量模式是否频繁.

高效用模式挖掘首先是在事务数据库的频繁项集挖掘中被提出[6]. 文献[18]通过定义一个统一的效用函数，将多个基于效用的度量合并到挖掘过程中，从而设计了一种统一高效用项集挖掘框架. 此外，考虑到增量与交互式挖掘的重要性，文献[19]提出了三种新的树结构有效地解决了增量和交互式高效用项集挖掘的问题. 空间高效用模式挖掘将空间模式挖掘和高效用项集挖掘进行了融合，不仅考虑到空间实例间的邻近关系，而且考虑了空间实例的效用值，以此挖掘邻近出现并且效用值高的co-location模式. 目前对空间高效用模式挖掘的研究主要包含特征带效用[7]和实例带效用[8-9]两种不同的挖掘方法. 文献[7]采用了类似于事务数据库中的方法，定义空间特征的外部效用和内部效用，并通过求取每个模式与整个数据库的效用占比来判断该模式是否是高效用的模式；文献[8]研究了实例带效用的高效用模式挖掘问题，定义了效用参与率和效用参与度等概念；文献[9]在文献[7]和[8]的基础上，定义和计算了每个特征在模式中的效用参与率，不仅解决了文献[7]中没有考虑特征效用之间差异的问题，而且解决了文献[8]中人为规定权重的问题. 其他高效用模式挖掘的研究，如文献[10]研究了空间高效用模式的增量挖掘问题；文献[20]提出了一种挖掘高效用序列模式的有效算法；文献[21]在考虑空间特征的效用属性的情况下，综合考虑了空间特征的时间属性，定义了空间实例的时间区间重叠和空间邻居的概念，提出了一种基于时间区间的高效用co-location模式挖掘的基本算法和剪枝算法.

k-近邻在空间数据中一般是指离目标实例最邻近的k个实例或邻居. 最初用在机器学习算法之一的k最近邻分类算法，思想是如果一个样本在特征空间中的k个最相邻的样本中的大多数属于某一个类别，则该样本也属于这个类别，并具有这个类别上样本的特性. 由于k-近邻思想简单又实用，越来越多的学者把k-近邻用到了数据挖掘和其他领域. 例如文献[22]和[23]将k-近邻的思想用于空间查询，在候选点集中查询和关键字最近的k个点. 文献[24]结合co-location模式挖掘和k-近邻提出了基于k-近邻特征的定性空间co-location模式挖掘算法. 文献[25]提出了面向不确定数据的k-近邻算法. 文献[26]则将模糊的k-近邻技术用于运动图像分类. 本文将k-近邻的思想运用到空间高效用co-location模式挖掘中，较好地解决了实例之间邻近关系定义的问题，同时，从实际应用需求出发，定义了核元素和核模式等概念，提出了一种新颖的k-近邻关系下的空间核模式挖掘框架.

3　相关定义和性质

本节首先给出挖掘框架中所使用到的定义，之后分析新概念具有的一些性质，除了定义1、定义2和定义3外，其余所有定义均由本文给出.

定义 1.　带效用的空间实例. 给定一个空间数据集S，其空间特征集为$F=\{f_1,f_2,\cdots,f_n\}$，空间实例集为$I=\{i_1,i_2,\cdots,i_m\}$，则空间特征$f_h(f_h\in F)$的第j个实例所带的效用v记为$f_h.j^v$，形式化表示为$u(f_h.j)=v$.

例如，图1是空间特征集{A,B,C,D}的带效用实例分布示例图，可以看到，$A.1^{10}$是特征A的第1个带效用实例，即$u(A.1)=10$.

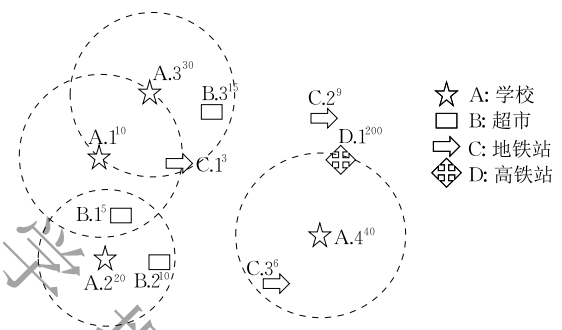

图1　一个带效用的空间实例分布示例图

定义 2.　空间实例的k-近邻实例集k-$NI(i_j)$(k-nearest instances set of a feature instance). 给定参数k，一个空间实例i_j的k-近邻实例集k-$NI(i_j)$是空间实例集I内与i_j最邻近的k个其他类型空间特征的实例(两两实例之间的邻近用欧几里得距离度量).

例如，在图1中，虚线圆表示的是特征A的4个实例的2-近邻实例集，于是，2-NI(A.1)={B.1,C.1}，2-NI(A.2)={B.2,B.1}，2-NI(A.3)={B.3,C.1}，2-NI(A.4)={C.3,D.1}.

定义 3.　空间实例的k-近邻特征集k-$NF(i_j)$(k-nearest feature types set of a feature instance). 给定一个空间实例i_j，则i_j的k-近邻特征集是k-$NI(i_j)$中的实例所属的特征集合，定义为

$$k\text{-}NF(i_j)=\{f_1,f_2,\cdots,f_x\mid\forall f_j\in k\text{-}NI(i_j).f\}.$$

例如，在图1中，$k=2$时，2-NF(A.1)={B,C}，2-NF(A.2)={B}，2-NF(A.3)={B,C}，

$2\text{-}NF(A.4)=\{C,D\}$.

定义 4. 空间特征的 k-近邻特征集 $k\text{-}NF(f_i)$ (k-nearest feature types set of a feature type)给定一个空间特征 f_i，f_i 的 k-近邻特征集等于该特征所有实例的 k-近邻特征集 $k\text{-}NF(i_j)$ 的并集，形式化定义如下：

$$k\text{-}NF(f_i)=\bigcup_{i_j\in ins(f_i)} k\text{-}NF(i_j).$$

同样地 k 取 2，由定义 4 可以得到：$2\text{-}NF(A)=\{2\text{-}NF(A.1)\cup 2\text{-}NF(A.2)\cup 2\text{-}NF(A.3)\cup 2\text{-}NF(A.4)\}=\{B,C,D\}$.

性质 1. 空间特征的 k-近邻特征集不满足对称性和自反性.

证明. 对称性(举反例). 如图 1 中，以特征 A、B 为例；设 $k=2$，已经求得 $2\text{-}NF(A)=\{B,C,D\}$. 特征 B 的实例为 $\{B.1,B.2,B.3\}$，$2\text{-}NI(B.1)=\{A.1,A.2\}$，$2\text{-}NI(B.2)=\{A.2,C.3\}$，$2\text{-}NI(B.3)=\{A.3,C.1\}$，则实例 B.1、B.2、B.3 的 2-近邻特征集为 $2\text{-}NF(B.1)=\{A\}$，$2\text{-}NF(B.2)=\{A,C\}$，$2\text{-}NF(B.3)=\{A,C\}$，所以特征 B 的 $2\text{-}NF(B)=\{2\text{-}NF(B.1)\cup 2\text{-}NF(B.2)\cup 2\text{-}NF(B.3)\}=\{A,C\}$. 特征 A 的 2-近邻特征集为 $\{B,C,D\}$，但是特征 B 的 2-近邻特征集为 $\{A,C\}$，故不满足对称性.

自反性. 根据实例的 k-近邻实例集的定义可知，空间实例 i_j 的 $k\text{-}NI(i_j)$ 中是不会出现和 i_j 属于同一特征的实例的，所以空间特征 f_i 不会出现在它的 k 近邻特征集中，故不满足自反性. 证毕.

定义 5. 核模式(core Pattern)和核元素(core element). 给定空间特征 f 和空间特征串 P，核模式 $\{f,P\}$ 表示特征 f 的 k-近邻内频繁地出现特征串 P 的实例，特征 f 被称为模式的核元素，核模式 $\{f,P\}$ 中特征总数($|P|+1$)称为核模式的阶.

根据特征的 k-近邻特征集的定义，可以知道给定核元素 f 会得到 f 的 k-近邻特征集 $k\text{-}NF(f)$，如核元素为 A，由定义 4 中的例子可知 $2\text{-}NF(A)=\{B,C,D\}$，所以可以根据 $2\text{-}NF(A)$ 进行组合得到 2 阶的候选核模式 $\{A,B\}$，$\{A,C\}$，$\{A,D\}$ 和 3 阶、4 阶的候选核模式 $\{A,BC\}$，$\{A,BD\}$，$\{A,CD\}$，$\{A,BCD\}$. 区别于传统的 co-location 模式，本文将新型的模式称为核模式，核模式中特征串的地位是平等的，如核模式 $\{A,BD\}$ 和 $\{A,DB\}$ 是同一个模式. 结合定义中的核模式与实例带效用的空间高效用 co-location 模式挖掘，本文提出了 k-近邻关系下的空间高效用核模式挖掘算法.

用一个小例子简单说明本文拟挖掘的高效用核模式在实际中的应用.

假设用户是一个创业者，想要在城市中的各大高校旁边以获取高收益为目标、开设酒店类或者餐饮类的服务项目，则可以以学校作为核元素挖掘高效用的核模式，当然，这里的高效用核模式与传统的高效用 co-location 模式不同，因为这个核模式本身并不关心学校的效用价值. 再接着思考，如果他又想看开设的酒店或餐馆周边哪些特征(如停车场、娱乐设施等)的组合具有高收益，则又可以以酒店或餐馆作为核元素进行进一步分析，如果采用传统的空间高效用 co-location 模式挖掘方法，则会将核元素的效用值一起考虑到挖掘到的模式中，这显然是不合适的. 那么对于提出的核模式，如何度量其效用高低呢？

首先本文给出核模式的参与实例集的定义，然后定义核模式中特征的效用参与率以及核模式的模式效用度，以此来度量核模式的效用高低.

定义 6. 参与实例集 $p\_instances$(set of participated instance). 给定核模式 $\{f,P\}$(f 为核元素，P 为特征串)，P 中特征 f_i 的参与实例集是核元素 f 的实例 $f.l$ 的 k-近邻实例集 $k\text{-}NI(f.l)$ 中，同时支持特征串 P 的特征 f_i 的实例的集合，即 $p\_instances(\{f,P\},f_i)=\{i_j\mid i_j.f=f_i\wedge i_j\in k\text{-}NI(f.l)\wedge P\subseteq k\text{-}NF(f.l)\}$，其中，$k\text{-}NF(f.l)$ 是核元素 f 的实例 $f.l$ 的 k-近邻特征集(见定义 3).

例如，对于图 1 中的核模式 $\{A,B\}$，核元素 A 实例的 $2\text{-}NI$ 中特征 B 的实例集为 $\{B.1,B.2,B.3\}$，所以核模式 $\{A,B\}$ 中特征 B 的参与实例集为 $\{B.1,B.2,B.3\}$，同理 $2\text{-}NI(A.4)$ 同时出现实例 C.3 和 D.1 支持特征 CD，故核模式 $\{A,CD\}$ 的 $p\_instances(\{A,CD\},C)=\{C.3\}$，$p\_instances(\{A,CD\},D)=\{D.1\}$.

定义 7. 特征的效用参与率 $FUPR$(Participation Ratio of Feature Utility). 给定核模式 $\{f,P\}$，则 P 中特征 f_i 的效用参与率定义为 f_i 在核模式 $\{f,P\}$ 参与实例集中的实例效用值之和除以特征 f_i 的所有实例效用值总和. 形式化定义如下：

$$FUPR(\{f,P\},f_i)=\frac{\sum_{i_j\in p\_instance(\{f,P\},f_i)} u(i_j)}{U(f_i)}.$$

$U(f_i)$ 表示特征 f_i 的总效用.

例如，设核元素为 A，根据定义 6 求得核模式 $\{A,CD\}$ 的参与实例集，特征 C 的带效用实例为 $\{C.1^3,C.2^9,C.3^6\}$，则特征 C 在核模式 $\{A,CD\}$ 中的特

征效用参与率为 $FUPR(\{A,CD\},C)=\frac{6}{3+9+6}=\frac{1}{3}$.

性质 2.　$0\leq FUPR(\{f,P\},f_i)\leq 1$.

证明.　根据参与实例集的定义可知，特征 f_i 的实例在核模式 $\{f,P\}$ 的参与实例集中出现的个数最多为该特征的所有实例个数，最少为 0 个，所以，$0\leq FUPR(\{f,P\},f_i)\leq 1$.　证毕.

特征效用参与率的计算是将不同特征的效用值做了归一化处理，解决了不同特征的效用不可相加的问题.

引理 1.　特征效用参与率的不增性：特征的效用参与率随着核模式阶数的增大而不增.

证明.　给定核模式 $\{f,P\}$ 和它的超模式 $\{f,P'\}(P\subseteq P')$，根据参与实例集的定义可知：$|p\_instances(\{f,P\},f_i)|\geq|p\_instances(\{f,P'\},f_i)|$，所以 $FUPR(\{f,P\},f_i)\geq FUPR(\{f,P'\},f_i)$.　证毕.

定义 8.　核模式效用度 $CPUI$（Utility Index of Core Pattern）. 核模式 $\{f,P\}$ 的效用度定义为 P 中所有特征的特征效用参与率的平均值，形式化表达为

$$CPUI(\{f,P\})=\frac{\sum_{f_i\in P}FUPR(\{f,P\},f_i)}{|P|}.$$

例如，对于核模式 $\{A,CD\}$，在定义 7 中已求得 $FUPR(\{A,CD\},C)=1/3\approx 0.33$，特征 D 的实例只有 D.1，$u(D.1)=200$，则 $FUPR(\{A,CD\},D)=1$，因此核模式 $\{A,CD\}$ 的效用度 $CPUI(\{A,CD\})=(0.33+1)/2=0.665$.

定义 9.　高效用核模式. 给定效用度阈值 ε，如果核模式 $\{f,P\}$ 的模式效用度 $CPUI(\{f,P\})\geq\varepsilon$，则称核模式 $\{f,P\}$ 为空间高效用 co-location 核模式（简称高效用核模式）.

例如，给定 $\varepsilon=0.5$，求得核模式 $\{A,CD\}$ 的模式效用度 $CPUI(\{A,CD\})=0.665>0.5$，所以 $\{A,CD\}$ 是一个高效用核模式.

核模式 $\{f,P\}$ 效用度的定义考虑了 P 中每个特征效用参与率的贡献，使得原来效用参与率较小的特征在模式中贡献小，效用参与率较大的特征在模式中的贡献较大，所以在一个不是高效用的核模式 $\{f,P\}$ 中加入一个效用参与率较高的特征后得到的新模式可能是高效用的. 因此，本文定义的核模式效用度 $CPUI$ 不能采用类先验原理的反单调性质[2]对候选模式进行剪枝，因为随着核模式阶数的增大 $CPUI$ 不一定单调递减，可能不减反增.

例如：如图 2 所示，圆圈内是特征 C 实例的 2-近邻实例集，根据前面的定义，以 C 为核元素的 2-近邻特征集 $2\text{-}NF(C)=\{A,B,D\}$，核模式 $\{C,A\}$ 的参与实例集为 $\{A.1\}$，$u(A.1)=0.2$，$\{C,A\}$ 的核模式效用度 $CPUI(\{C,A\})=0.2$，设效用度阈值为 0.3，可知 $\{C,A\}$ 不是高效用的核模式，但对于它的超模式 $\{C,AB\}$，其参与实例集分别为 $\{A.1\}$ 和 $\{B.1,B.2\}$，则 $\{C,AB\}$ 的核模式效用度 $CPUI(\{C,AB\})=\left(\frac{u(B.1)+u(B.2)}{u(B.1)+u(B.2)+u(B.3)}+0.2\right)\Big/2=0.52$，所以核模式 $\{C,A\}$ 的超模式 $\{C,AB\}$ 是一个高效用核模式，可见对一个核模式 $\{f,P\}$，其超模式 $\{f,P'\}$ 的模式效用度会存在：$CPUI(\{f,P'\})>CPUI(\{f,P\})$.

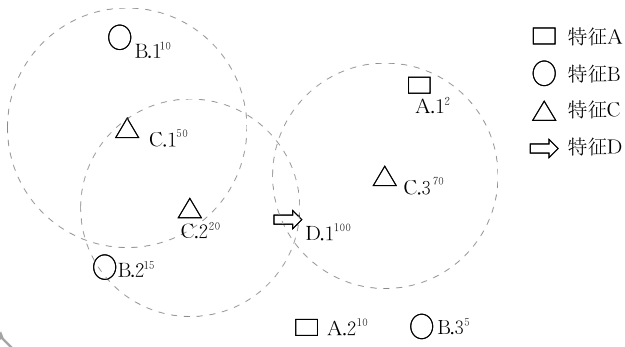

图 2　核模式 $CPUI$ 小于其超模式 $CPUI$ 的空间实例图例

4　挖掘算法

本节设计了一个基本的高效用核模式挖掘算法和一个优化算法.

4.1　基本算法 BA（Basic Algorithm）

算法 BA 挖掘思路. 首先计算用户给定的核特征集 cF 中每个核元素 f 实例的 k-近邻实例集 $k\text{-}NI$，然后得到 f 的 k-近邻特征集 $k\text{-}NF(f)$，进而形成核元素 f 的候选核模式，但是当 cF 中特征较多或者核元素的实例个数较大的时候，用传统的基于距离的计算方法时间复杂度较高. 为了提高算法效率，本文采用了一种基于网格的计算方法，能快速求出实例的 $k\text{-}NI$. 实例 i_j 的邻近网格是其所在网格的周围（上下左右）八个网格.

首先将空间实例映射到网格中（假设实例分布范围 1000×1000），然后给定网格边长 d（这里取 20），即将网格划分为 50×50 个小网格，如图 3 为部分网格示例，假设核元素为 A，以 A 的实例 A.1 为例说明网格法求 $k\text{-}NI$ 的计算过程：以 A.1 为圆心，半径 $R=d$ 画圆（虚拟圆），求出实例 A.1 所在网格及邻近网格中和 A.1 特征不同的实例与 A.1 的距离，如

果距离小于等于 d 的实例个数大于等于 k 个，即在 A.1 的所在网格及邻近网格中取 A.1 的 k-NI 即可；如果求出小于等于 d 的实例个数不到 k 个，将圆的半径＋d 再找出阴影部分网格的邻近网格循环计算.

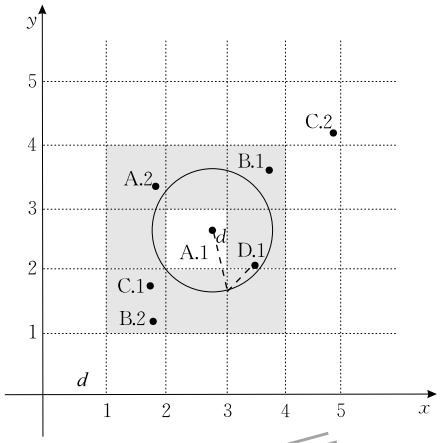

图 3 网格法计算示例

网格边长 d 的选取. 网格边长一般根据空间实例的分布情况和 k 的大小来选取. 假设空间范围为 1000×1000，空间实例的个数为 20 000，$k=5$，网格的边长为 d，那么平均一个网格有 $\frac{20000}{1000\times1000/d\times d}$ 个实例，由于 $k=5$，那么一个格子的实例个数在个位数左右即可快速获得实例的 k-近邻实例集，如 $d=20$，平均一个网格有 20 000/2500＝8 个实例.

根据提出的核模式及其特性，本文提出一种“顺序树”的方法快速收集所有的候选核模式.

顺序树的建立. 以核元素 A 来说明顺序树的构成和候选核模式的收集过程. 假设 A 的 k-近邻特征集 k-NF(A)＝{C,D,B,E}，将其按照字典顺序排序为{B,C,D,E}，以核元素 A 为根节点，k-NF(A)按照字典顺序作为 A 的孩子结点(第二层)，接着将第二层结点依次作为当前核元素，以它字典排序后的特征作为孩子结点循环…得到如图 4 所示的顺序树结构.

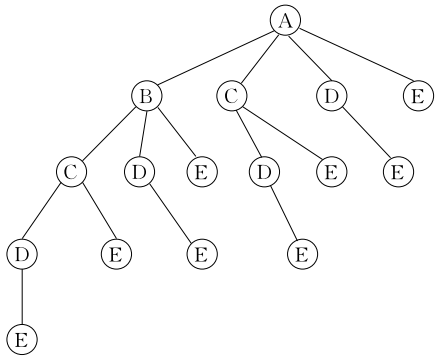

图 4 顺序树方法收集候选核模式示例

候选核模式的收集. 层次遍历，从根结点 A 依次向下层次遍历每个结点并进行组合，到第二层形成二阶候选，到第三层形成三阶候选…. 表 1 是从图 4 所示顺序树收集到的所有候选核模式.

表 1 图 4 中顺序树收集到的候选核模式

层数	候选核模式
2	{A,B},{A,C},{A,D},{A,E}
3	{A,BC},{A,BD},{A,BE},{A,CD},{A,CE},{A,DE}
4	{A,BCD},{A,BCE},{A,BDE},{A,CDE}
5	{A,BCDE}

算法 1 给出了挖掘所有高效用核模式的一个基本算法，简称算法 BA. BA 对核特征集 cF 中每个核元素，依次采用“候选-测试”机制逐阶挖掘高效用核模式. 步骤 1 对一个核元素 f，基于网格方法计算核元素每个实例的 k-近邻实例集 k-NI，基于定义 3 和 4，求出核元素 f 的 k-近邻特征集 k-$NF(f)$，接着用顺序树的方法收集所有的候选核模式{f,P}. 步骤 2 从 2 阶开始，逐阶计算候选模式的效用度，如果效用度值不小于 ε，则将其并入高效用核模式集 CP，直到没有任何高效用核模式生成为止，算法最终返回所有的高效用核模式集 CP.

算法 1. 基本算法 BA.

输入：空间特征集 F，带效用的空间实例集 I，参数 k，核元素集 cF，效用度阈值 ε

输出：高效用核模式集 CP

步骤：

BEGIN

1. FOR EACH $f\in cF$

1.1 FOR EACH $i_j\in f$

1.1.1 基于网格的方法计算实例 i_j 的 k-$NI(i_j)$；

1.2 通过 k-$NI(i_j)$求出核元素 f 的 k-$NF(f)$；

1.3 根据 k-$NF(f)$特征集采用顺序树的方法收集所有的候选核模式{f,P}；

2. FOR EACH {f,P} //由低阶→高阶

2.1 生成{f,P}中每个特征的参与实例集；

2.2 计算{f,P}中每个特征的特征效用参与率 $FUPR$；

2.3 计算{f,P}的模式效用度 $CPUI$，如果其效用度值不小于 ε，{f,P}→CP；

3. 输出高效用的核模式集 CP

END

4.2 优化算法 OA(Optimized Algorithm)

由于基础算法采用了“候选-测试”的机制，算法效率不高，并且定义的核模式效用度 $CPUI$ 不满足反单调性质，不能采用类先验原理的方法进行剪枝，所以本小节提出 4 个剪枝优化策略有效地提高算法

效率，将经过剪枝优化后的算法记为 OA.

剪枝 1. 给定 m 阶的核模式$\{f,P\}$，若它的所有参与实例集为空，则$\{f,P\}$以及它的所有超集$\{f,P'\}$一定不是高效用核模式.

证明. 根据定义 6 可知，核模式$\{f,P\}$的参与实例集是核元素 f 实例的 k-近邻实例集 $k\text{-}NI(i_j)$ $(i_j.f=f)$中支持特征串 P 的实例的集合，所以在低阶的核模式中若没有实例支持 P，那么它的所有超集$\{f,P'\}$的参与实例集中肯定也不会有支持 P' 的实例，即可剪枝. 证毕.

例如，图 5 中圆圈里表示的是特征 A 的 2-近邻实例集，$2\text{-}NI(\text{A.1})=\{\text{B.1},\text{C.1}\}$，$2\text{-}NI(\text{A.2})=\{\text{B.2},\text{D.1}\}$，所以 $2\text{-}NF(\text{A})=\{\text{B},\text{C},\text{D}\}$，候选核模式有{A,B}，{A,C}{A,D}，{A,BC}，{A,BD}，{A,CD}，{A,BCD}. 由参与实例集的定义可知核元素 A 的 2-近邻实例集中没有同时出现支持 CD 的实例，所以核模式{A,CD}以及超集{A,BCD}可被剪枝.

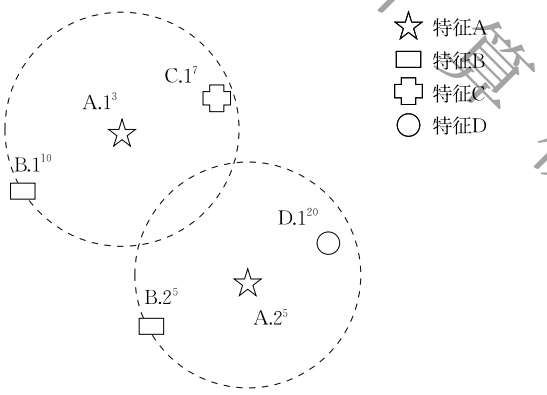

图 5　剪枝 1 空间实例示例

剪枝 2. 给定核模式$\{f,P\}(|P|>1)$，当 P 中所有特征在对应二阶核模式中的特征效用参与率 $FUPR$ 都小于效用度阈值 ε 时，$\{f,P\}$一定不是一个高效用核模式.

证明. 由引理 1 特征效用参与率的不增性可得：因为 $\forall f_i\in P$，$FUPR(\{f,f_i\},f_i)<\varepsilon$，所以 $CPUI(\{f,P\})=\dfrac{\sum_{f_i\in P}FUPR(\{f,P\},f_i)}{|P|}<\varepsilon$. 证毕.

剪枝 3. 给定核模式$\{f,P\}(|P|>1)$，当 P 中所有特征在对应二阶核模式中的特征效用参与率的平均效用参与率$<\varepsilon$时，$\{f,P\}$一定不是高效用核模式.

证明. 根据核模式参与实例集的定义，特征 f_i 的实例在二阶核模式中出现的个数一定大于等于其在高阶的核模式$\{f,P\}$中出现的实例个数，即 $|p\_instances(\{f,f_i\},f_i)|\geq|p\_instances(\{f,P\},f_i)|$，所以当 $\dfrac{\sum_{f_i\in P}FUPR(\{f,f_i\},f_i)}{|P|}<\varepsilon$ 时，核模式$\{f,P\}$的模式效用度 $CPUI(\{f,P\})=\dfrac{\sum_{f_i\in P}FUPR(\{f,P\},f_i)}{|P|}\leq\dfrac{\sum_{f_i\in P}FUPR(\{f,f_i\},f_i)}{|P|}<\varepsilon$.

证毕.

剪枝 4. 给定 $m(m>1)$阶的核模式$\{f,P\}$，如果它的模式效用度 $CPUI(\{f,P\})<\varepsilon$，则对$\{f,P\}$及其超模式$\{f,P'\}$求差集得到的特征集 $F'(F'\subseteq F)$，取 F'中所有特征 f_i在二阶核模式$\{f,f_i\}$中的特征效用参与率的平均值与$\{f,P\}$的模式效用度求平均，即 $\left(CPUI(\{f,P\})+\dfrac{\sum_{f_i\in F'}FUPR(\{f,f_i\},f_i)}{|F'|}\right)\Big/2<\varepsilon$，则超模式$\{f,P'\}$一定不是高效用核模式.

证明. 对 $m(m>1)$阶的$\{f,P\}$和它的超集$\{f,P'\}$$(P\subseteq P')$，由引理 1 特征效用参与率的不增性有：$FUPR(\{f,P'\},f_j)\leq FUPR(\{f,P\},f_j)<\varepsilon$，$P$ 与 P'求差集得 $F'\{F'|F'\subseteq P'\wedge F'\not\subset P\}$，$F'$中的特征 f_i在超模式$\{f,P'\}$中的特征效用参与率小于等于 f_i在二阶核模式$\{f,f_i\}$中的特征效用参与率，故 $\dfrac{\sum_{f_i\in F'\wedge F'\subseteq P'}FUPR(\{f,P'\},f_i)}{|F'|}\leq\dfrac{\sum_{f_i\in F'}FUPR(\{f,f_i\},f_i)}{|F'|}$，又因$\{f,P\}$的 $CPUI(\{f,P\})<\varepsilon$，$|P'|=|P|+|F'|$，所以$\{f,P'\}$的模式效用度 $CPUI(\{f,P'\})=\dfrac{\sum_{f_j\in P'}FUPR(\{f,P'\},f_j)}{|P'|}\leq\left(CPUI(\{f,P\})+\dfrac{\sum_{f_i\in F'}FUPR(\{f,f_i\},f_i)}{|F'|}\right)\Big/2<\varepsilon$.

证毕.

算法 2 是基于剪枝 1～剪枝 4 设计的一个优化算法，简称算法 OA. 步骤 1 生成候选核模式$\{f,P\}$与算法 1 相同. 为应用剪枝 2～剪枝 4，步骤 2 计算二阶候选核模式的效用度，并用剪枝 2 和剪枝 3 对候选进行剪枝. 步骤 3 中，计算候选的参与实例、效用参与度，并用剪枝 1 和剪枝 4 对候选核模式进行剪枝.

算法 2. 优化算法 OA.

输入：空间特征集 F，带效用实例集 I，距离阈值 d，核元素集 cF，效用度阈值 ε

输出：高效用的核模式 CP

步骤：

BEGIN

1. FOR EACH $f \in cF$

1.1 求出核元素 f 实例的 k-NI 和 f 的 k-$NF(f)$；

1.2. 根据 k-$NF(f)$生成候选核模式$\{f,P\}$；

2. FOR EACH $\{f,f_i\}(f_i \in P)$ //二阶候选

2.1 计算$CPUI(\{f,f_i\})$并记录，若$CPUI(\{f,f_i\}) \geq \varepsilon$，$\{f,f_i\} \rightarrow CP$；

2.2 用剪枝 2 和剪枝 3 对候选进行剪枝；

3. FOR EACH $\{f,P\}(|P|>1)$ //三阶以上候选

3.1 生成核模式$\{f,P\}$的参与实例集，若为 null，用剪枝 1 剪枝；

3.2 计算$\{f,P\}$的特征效用参与率 $FUPR$；

3.3 计算$\{f,P\}$的模式效用度 $CPUI(\{f,P\})$；

若 $CPUI(\{f,P\}) \geq \varepsilon$，$\{f,P\} \rightarrow CP$；否则用剪枝 4 对候选进行剪枝；

4. 输出高效用的核模式集 CP

END

4.3 算法分析

本小节主要分析算法 1 和算法 2 的复杂度并讨论算法的完备性和正确性.

4.3.1 复杂性分析

时间复杂度. 假设核特征集中核实例的数目是 N_1，其他非核特征的实例总数是 N_2；核特征数 $|cF|=M_1$，非核特征数是 m. 在算法 1 中，步骤 1 计算核实例的 k-近邻，使用网格方法，最坏情况下的时间复杂度约 $9\times k_1 \times N_1 = O(k_1 \times N_1)$（$k_1$ 为网格中的平均实例数）；而应用顺序树方法生成候选核模式的复杂度最坏情况下是 $O(M_1 \times 2^m)$，其中 m 是核元素的 k-近邻特征集的平均数. 可以注意到的是，$O(M_1 \times 2^m)$也是候选核模式的数目. 步骤 2 中，假设候选模式的平均长度为 s_1，核特征的平均实例数是 s_2，那么，计算每个候选模式的复杂度约为 $O(s_1 \times s_2)$，于是，步骤 2 计算复杂度是 $O(M_1 \times 2^m \times s_1 \times s_2)$. 显然，候选模式数目是制约算法效率的关键. 在算法 2 中，经过剪枝 1 到剪枝 4 的剪枝，算法效率将得到有效提升. 当然，算法效率与特征数、实例数以及实例分布等有极大的关系.

空间复杂度. 两个算法的空间耗费主要来自两个方面：(1) 保存每个核元素的 k-近邻实例集和 k-近邻特征集，k-近邻特征集的数目一般小于 k-近邻实例集的数目，所以，这部分的空间复杂度约为 $O(k \times N_1)$，其中 N_1 是核特征集中核实例的数目，这个复杂度不会超过传统 co-location 模式挖掘算法中保存的邻近关系数；(2) 保存候选模式及高效用核模式，空间消耗为 $O(s_1 \times M_1 \times 2^m)$，其中 s_1 是候选模式的平均长度，M_1 是核特征数 $|cF|$，m 是核元素的 k-近邻特征集的平均数. 顺序树知识候选生成的一种方式，实际计算中没有存储树结构. 另外，算法 1 可以不用提前产生所有候选，可以减少这部分内存耗费. 而算法 2 除了候选需要产生，还需要存储所有二阶非高效用核模式，内存占用比算法 1 稍多些. 但是，与传统 co-location 模式挖掘算法相比，两个算法没有占用额外的存储空间.

4.3.2 完备性和正确性

完备性表示算法能够找到所有的高效用核模式，正确性意味着算法找到的每个高效用核模式的效用度不小于效用度阈值 ε.

完备性. 算法 1 和算法 2 是完备的. 算法 1 基于核特征的 k-近邻特征集产生候选核模式，不会遗漏任何与核元素有邻近关系的核模式. 因此，算法 1 一定能搜索到所有的高效用核模式. 而算法 2 中应用的剪枝 1～剪枝 4 剪去的候选核模式一定不是高效用的（已在剪枝 1～剪枝 4 中证明），因此，两个算法能够找到所有的高效用核模式.

正确性. 算法 1 和算法 2 都是正确的. 算法的正确性由两方面保证：(1) 核模式效用度的计算是正确的，算法中严格按定义 7 和定义 8 的公式进行了计算，计算结果的正确性有保证；(2) 算法 1 中步骤 2.3 保证了 CP 中每个高效用核模式的效用度不小于效用度阈值 ε，算法 2 中步骤 2.1 和 3.3 保证了 CP 中每个高效用核模式的效用度不小于效用度阈值 ε.

5 实验评估

本节在真实和模拟数据集上进行了广泛的实验验证本文所提挖掘算法的实用性和先进性，并分析了不同参数阈值和不同数据密度对算法效率的影响以及优化算法的效果. 实验环境为 Win10 系统，16 GB 内存，CPU 为 Intel Core i7，编程语言为 Java.

5.1 算法的实用性分析

本小节在一个真实数据上与文献[9]提出的

PUI 算法进行对比来分析本文所提算法(记为 CPUI)的挖掘结果的实用性和优越性，PUI 是空间高效用 co-location 模式挖掘方面近两年比较有代表性的算法，该算法较好地解决了文献[7](特征带效用)和文献[8](实例带效用)存在的一些问题，并且其挖掘结果是优于这两者的. 所选择的真实数据集为北京市的一些 POI(兴趣点)数据，包含空间特征 16 个，空间实例为 23025 个，图 6 展示了其空间实例的分布情况，表 2 详细解释了各个空间特征代表的实际意义和对应的实例个数.

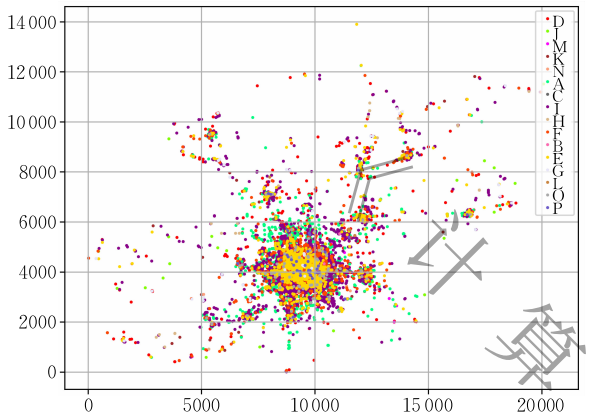

图 6　北京市 POI(兴趣点)数据分布图

表 2　真实数据集的特征实际意义及实例个数

顺序号	特征	实际含义	实例个数
1	A	中餐厅	4685
2	B	西餐厅	28
3	C	咖啡屋	1575
4	D	酒店	3439
5	E	招待所	1560
6	F	花园	953
7	G	国家级景点	53
8	H	省级景点	29
9	I	停车场	8343
10	J	市区公交站	61
11	K	火车站	73
12	L	机场	10
13	M	服饰店	1994
14	N	宠物商店	953
15	O	家电超市	81
16	P	体育用品超市	47

如图 6 以及表 2 所示，该数据的实例分布包含了密集和稀疏的区域，且各个特征的实例数相差较大，包含了一些实例数较少的稀有特征. 在这样一个空间分布和实例数量上都不均匀的数据集上，本小节从将普通特征 C(咖啡屋)和稀有特征 G、K(国家级景点、火车站)作为核元素，并且在效用度阈值相同，k 远小于 d 的情况下对挖掘结果进行对比和分析.

实验结果为表 3 和表 4. 表 3 是核元素集为{C}的结果，表 4 是核元素集为稀有特征{G,K}的结果，PUI 算法列出带有核元素的高效用模式，具体的实验参数皆在表中.

表 3　C(咖啡屋)作为核元素的挖掘结果比较

	CPUI 算法	PUI 算法
距离阈值 d	×	20 m
参数 k	10	×
效用度阈值	0.2	0.2
核元素集	{**C**}	×
结果对比	{C,D}	CD
	{C,L}	CL
	{C,G}	×
	×	CN
	{C,IM}	CIM
	{C,AI}	CAI
	{C,BD}	×
	{C,DIL}	×
	…	…

表 4　核元素为稀有特征{G,K}的挖掘结果比较

	CPUI 算法	PUI 算法
距离阈值 d	×	20m
参数 k	10	×
效用度阈值	0.1	0.1
核元素集	{**G,K**}	×
结果对比	{G,H}	×
	×	GC
	{G,K}	×
	{K,J}	×
	{K,HI}	KI
	…	…

实验结果分析：

首先从表 3 可以看出，将普通特征作为核元素时，PUI 算法能挖掘出的高效用模式 CPUI 算法基本也能得到，并且本文的挖掘算法能挖掘出{C,BD}({咖啡屋,(西餐厅,酒店)})、{C,DIL}({咖啡屋,(酒店,停车场,机场)})等 PUI 算法挖掘不到但很有应用价值(如商业决策)的模式，而对于 PUI 得到的模式 CN({咖啡屋,宠物店})，现实生活中一般是不能将宠物带入餐饮店的，得到这个模式的原因正是将咖啡屋的效用值一起计算到模式中，得到的高效用模式并不一定符合实际应用.

再来看表 4：以稀有特征作为核元素时，由于稀有特征的实例在整个数据库中占比比较小，所以效用度阈值设为 0.1. 从结果可以看出，CPUI 算法对稀有特征的处理结果更好，能得到{G,H}({国家级景点,省级景点})、{K,J}({火车站,市公交站})等 PUI 算法遗失或没有挖掘出的十分有意义的带稀有特征的高效用模式，究其原因，是因为传统的空间高

效用 co-location 模式(PUI)在度量实例之间的邻近关系的时候给定的是单一的距离阈值,对不均匀的数据集处理效果不好(本实验真实数据集就是不均匀的),而本文的算法以一种比较"自然"的方法(k-近邻)来度量实例之间的邻近关系,对均匀/不均匀的数据集都有良好的可伸缩性. 同时,将国家级景点作为参照物(核元素)时,对于决策者来说是不考虑景点本身收益的,所以 PUI 算法得到的模式 GC({国家级景点,咖啡屋})类似于表 3 中的 CN,也是将核元素的效用值一起计算到模式中,而对于 CPUI 算法,咖啡屋出现在酒店、西餐厅、机场等(表 3)旁边的效用值更多,{G,H}{国家级景点,省级景点}等才是更有意义的高效用模式.

综上所述,本文的挖掘算法在实际中具有较高的应用价值.

5.2 算法的性能以及可扩展性分析

本小节主要是在真实数据(北京市 POI 数据)以及模拟数据上验证不同数据密度和不同参数对挖掘算法的影响以及与同类算法的性能比较,由于本文采用的是实例带效用的策略,所以对比算法也是目前现有的空间实例带效用的算法,分别为文献[7]的 UPI 优化剪枝算法(记为 UPI)以及文献[8]的 PUI 最大特征效用参与率剪枝算法(记为 PUI),同时对本文所提算法进行可扩展性分析. 为了便于区分比较,本文的基础算法记为 CPUI_BA,优化算法记为 CPUI_OA. 模拟数据的合成参照文献[2]中的方法,分别在不同的空间范围内生成了 5 组不同特性的模拟数据集,每个数据都包含了特征数,实例个数,实例效用范围,实例坐标范围四个属性,详细信息如表 5 所示.

表 5 合成数据相关信息

合成数据	特征数	实例个数	实例效用范围	实例坐标范围
Data_1	20	20000	[1,1000]	(2000,2000)
Data_2				(300,300)
Data_3		[20000,100000]		
Data_4	[10,25]	20000		(10000,10000)
Data_3		[10000,25000]		

首先,本文在 2000×2000、300×300 范围内分别生成了 Data_1 和 Data_2,这两个数据集的空间特征数都为 20,分别代表了不同密度的数据分布情况. 接着,Data_3、Data_4 和 Data_5 的实例坐标范围为 10 000×10 000. Data_3 的空间特征数为 20,而实例数在[20 000,100 000]范围内可变,最大实例数为 100 000. Data_4 和 Data_5 的特征数都是在[10,25]范围内可变的,而 Data_4 的实例数固定为 20 000,Data_5 的实例数的可变范围为[10 000,25 000]. Data_3、Data_4、Data_5 三个数据集,特征数和实例数可以改变,便于对算法进行定量和可扩展性分析. 以上所有的数据集中实例的效用值都大于 1 且小于 1000,所以实例的效用范围为[1,1000].

由于本文提出的新概念和新模式有别于传统 co-location 模式的相关概念和模式,所以与类似算法 UPI 和 PUI 只能在参数可控和类似范围里进行性能比较. 比如 k 的取值类似于传统的 co-location 模式中团的大小,给定 k 值后会得到实例的 k-近邻实例集,而 UPI 和 PUI 的距离阈值 d 应当随数据范围的大小而变化,数据范围分布较大的 d 应适当取得大一点,否则容易不能形成团.

参数 k 和距离阈值 d 变化对算法的影响:

本次选取的合成数据集为 Data_1,真实数据集为 5.1 节中北京的 POI(空间范围为 14 000×23 000),实验结果为图 7 和图 8. 参数设置:效用度阈值均为 0.3,CPUI 核元素个数为 1,k 的取值为 4～12,根据空间范围 UPI 和 PUI 距离阈值 d 为 40～120.

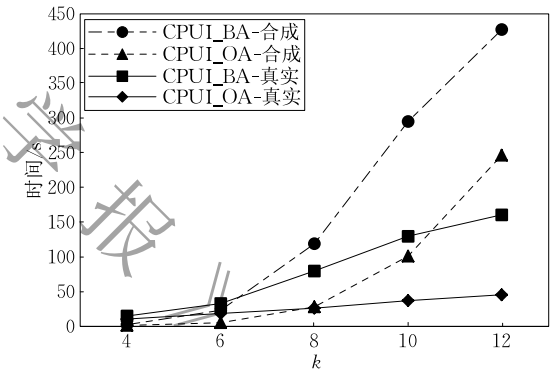

图 7 k 值对 CPUI 算法的影响

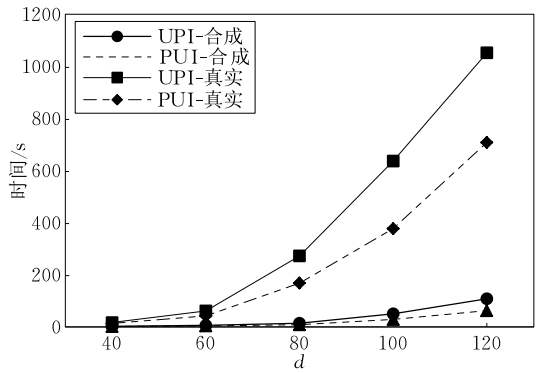

图 8 距离阈值 d 对 UPI 和 PUI 算法的影响

由实验结果可知:在真实数据和合成数据上,随着 k 值和距离阈值 d 的增加,UPI、PUI 和 CPUI 算法的时间复杂度依次上升,这是因为距离阈值 d 的

增加会使 UPI 和 PUI 算法得到更多实例的邻近关系，形成的团和模式数量增加，算法的运行时间上升；而 k 的取值决定了参与实例集的大小，类比于传统 co-location 模式挖掘中的团，k 越大时，核元素的 k-近邻实例集将会包含更多空间特征的实例，继而形成大量的候选核模式增加运行成本.

CPUI的表现：可以看到，本文优化算法CPUI_OA在整体上是优于 UPI 和 PUI 的，特别是在真实数据集上的表现，UPI 和 PUI 的时间花费分别到了大约 10 min 和 6 min，而 CPUI_OA 在 50 s 左右，所以本文所提的挖掘算法更加适用于实际应用，并且无论 k 取何值，CPUI_OA 算法的表现都优于 CPUI_BA 算法，说明了优化算法 CPUI_OA 的先进性和高效性.

不同数据密度和效用度阈值对算法的影响：

该实验选择的实验数据集为 Data_1 和 Data_2，Data_1 的空间范围为 2000×2000，比较稀疏（图中用 x 表示）；Data_2 的空间范围 300×300 的比较稠密（图中用 m 表示）. 参数设置：效用度阈值均从 0.1～0.6 变化，CPUI 的核元素个数为 1，$k=6$；由于空间范围较小，UPI 和 PUI 距离阈值 d 设为 20. 为了便于观察 CPUI 算法与 PUI 和 UPI 算法对不同数据密度的表现，将实验结果分为图 9 和图 10.

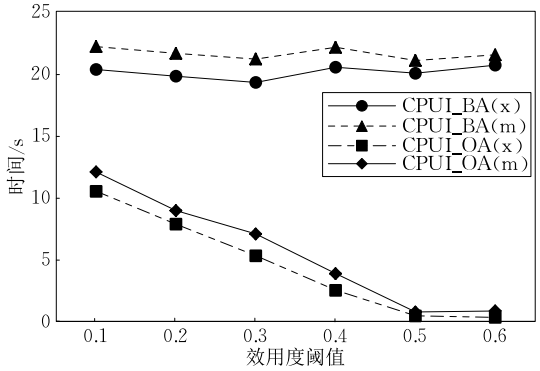

图 9　不同数据密度和效用度阈值对 CPUI 的影响

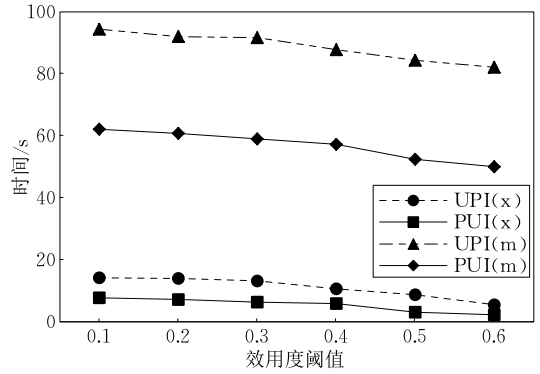

图 10　不同数据密度和效用度阈值对 UPI 和 PUI 的影响

不同数据密度的表现：在真实数据实验时，可以看到本文的算法对不均匀的数据集具有良好的表现，这一点在本小节同样得到了验证，在图 9 和图 10 的实验中本文采用了两组密度不同的数据进行了对比试验，从实验结果可以看出：UPI 和 PUI 算法在距离阈值不变、效用阈值变化时，对不同密度数据的处理结果有很大差异，这是因为两者在数据密度较稠密（300×300）的空间范围会得到大量的空间实例的邻近关系，而在数据密度较稀疏的空间范围（2000×2000）得到邻近关系又较少，所以也说明了 UPI 和 PUI 算法的不稳定性. CPUI 的表现：在图 9 中，无论是 CPUI_BA 算法还是 CPUI_OA 算法，对分布比较密集和稀疏的数据其运行时间基本一致，说明了本文所提算法对不同密度的数据表现好，具有较好的稳定性和可伸缩性.

不同效用度阈值的表现：（CPUI 的表现）从图 9 中可以看出，当效用度阈值从 0.1 逐步增加到 0.6 的时候，CPUI_BA 算法运行时间基本不变，而 CPUI_OA 算法的运行时间逐渐下降直到接近 0 点，这是因为 CPUI_BA 算法由“候选-测试”的机制来得到高效用的核模式，算法时间不受效用度阈值影响，而优化算法随效用度阈值的增加剪枝效果越加明显，并且当效用度阈值达到一个峰值（图中为 0.5）即所有二阶核模式的模式效用度都小于 ε 的时候，CPUI_OA 算法终止；相比于图 10 中的结果，可以发现随着效用度阈值的增加，UPI 和 PUI 算法都有一定的剪枝效果，但相比之下 CPUI_OA 的优化剪枝效果更加明显.

实例数目变化对算法的影响：

本次实验采用的数据集为 Data_3，实例数目从 20 000～100 000 变化. 参数设置：效用度阈值均设为为 0.2，CPUI 核元素个数为 1，$k=6$；UPI 和 PUI 距离阈值 d 根据空间范围设为 90.

由图 11 可知，当实例数目从 20 000 增加到

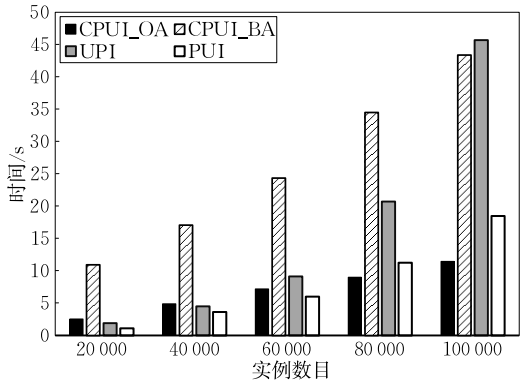

图 11　实例数对算法的影响

100 000 时，几个算法的运行时间逐步增加，这是因为实例数增加时 CPUI 中核元素的实例数也会增加，会得到更多的 k-近邻实例集和候选核模式，而 UPI 和 PUI 算法随着实例数增加会使数据密度更加稠密，邻近关系增多，UPI 算法随着实例到十万左右时优化性能较差，运行时间急剧上升.

CPUI 的表现：从图 11 可知：CPUI_OA 算法的运行时间同 PUI 优化剪枝策略的效果基本保持一个相当水平并且高于 PUI 算法；同时在同等的 k 值下由于 CPUI_OA 算法良好的剪枝效果其运行时间还不到基础算法的一半.

特征数目对算法的影响：

根据以往 co-location 模式挖掘的经验，特征的数量通常会影响挖掘算法的效率. 图 12 的实验选择了数据集 Data_4，在空间实例数不变的情况下，将特征的数量由 10 增加到 25. 参数设置：效用度阈值为 0.3，CPUI 核元素个数为 1，k 取 5. 由于空间范围为 20 000×20 000 较大，所以 UPI 和 PUI 距离阈值 d 设为 250.

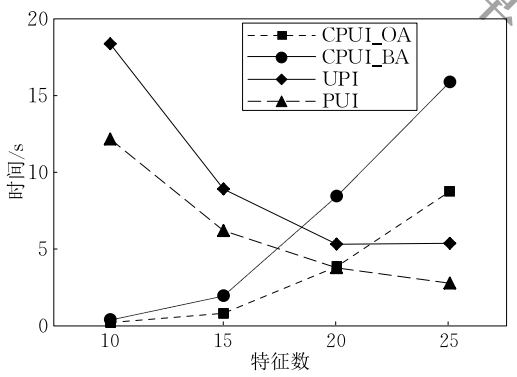

图 12 特征数对算法的影响

实验结果表明，随着特征数的增加，UPI 和 PUI 的时间复杂度依次降低，这是因为 UPI 和 PUI 算法在空间范围、距离阈值和实例数不变的情况下，单纯的增加特征的数目实际上是在减少每个特征所拥有的实例数目，导致支持形成候选模式的表实例数量减少，同时因为增加别的特征的实例将原本邻近的有些实例变得不再邻近，所以就减小了计算量，增加剪枝性能. 而 CPUI_BA 和 CPUI_OA 的执行时间都在增加. 这是因为 k-近邻的特性，增加了特征后，会计算更多特征的 k-近邻实例集，得到的候选也增多，将得到核模式的类型增多，比较符合实际逻辑.

特征数和实例数变化：为了保证和验证所提的算法有一定的可扩展性，图 13 的实验中同时改变了特征数和实例数，所选取的模拟数据为 Data_5. 其他参数设置和图 12 一样. 实验不同之处在于：增加特征的同时由于增加了每个特征的实例数目，不会导致 UPI 和 PUI 算法原来的邻近关系失效，而且会增加模式的多样性，使得算法的挖掘负担加重. 同时可以看到 CPUI_OA 的执行效率同图 12 中一样始终高于 CPUI_BA. 这也说明，随着数据的改变，挖掘算法具有一定的稳定性和可扩展性.

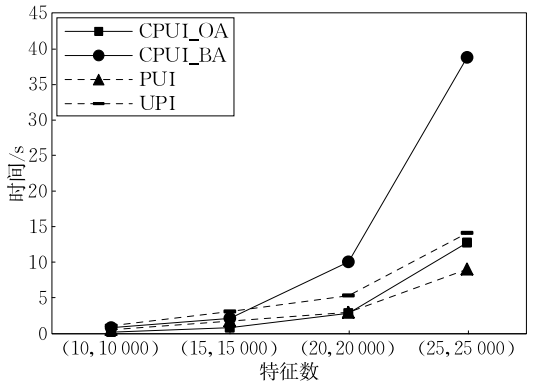

图 13 特征数和实例数对算法的影响

核元素个数对算法的影响：

最后是改变核元素的个数，增加本文所提算法的可扩展性分析. 图 14 分别在真实数据集和合成数据集上比较了核元素个数对本文所提算法效率的影响. 合成数据集为 Data_1，k 取值为 6，效用度阈值为 0.4，核元素个数的取值范围为[4,12]. 由图 14 可知，随着核元素个数的增加，CPUI_BA 和 CPUI_OA 的执行时间也随着增加，这是因为由于核元素个数的增加，候选模式的类型和数量必然是要增加的，从而使得更多的模式需要被计算与判断. 此外，从实验结果来看，无论是真实数据集还是合成数据集，CPUI_OA 都表现出了较为优异的执行效率.

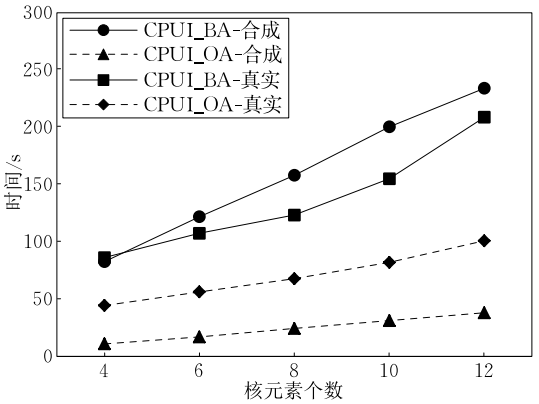

图 14 核元素个数对 CPUI 算法的影响

6 结束语

传统的空间高效用 co-location 模式挖掘和经典的 co-location 模式挖掘都存在着同样的缺陷，那就是人为地用单一距离阈值来度量空间实例之间的邻近关系，并且算法对不均匀的数据表现不好. 对此本文将 k-近邻关系引进空间高效用 co-location 模式挖掘中，提出了核元素和核模式的概念，设计了新的空间高效用核模式挖掘算法，由于核模式的效用度度量不满足反单调性质，提出的基本算法耗时较长，所以本文提出了 4 个剪枝策略对基本算法进行了优化. 在真实数据集和合成数据集上与现有的实例带效用的同类算法 UPI 和 PUI 进行对比实验，验证了本文所提挖掘算法的实用性、先进性、可伸缩性以及优化算法的高效性. 在今后的工作中将会在本文的基础上，同时考虑多个核元素的情况，并结合模糊集理论，研究基于模糊技术的 k-近邻高效用核模式挖掘的问题.

参 考 文 献

[1] Yao X, Chen L, Peng L, Chi T. A co-location pattern-mining algorithm with a density-weighted distance thresholding consideration. Information Sciences, 2017, 396: 144-161

[2] Huang Y, Shekhar S, Xiong H. Discovering co-location patterns from spatial data sets: A general approach. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(12): 1472-1485

[3] Lu J, Wang L, Fang Y, Zhao J. Mining strong symbiotic patterns hidden in spatial prevalent co-location patterns. Knowledge-Based Systems, 2018, 146: 190-202

[4] Yu W. Spatial co-location pattern mining for location-based services in road networks. Expert Systems with Applications, 2016, 46(Mar.): 324-335

[5] Yao X, Jiang X, Wang D, et al. Efficiently mining maximal co-locations in a spatial continuous field under directed road networks. Information Sciences, 2021, 542: 357-379

[6] Yao H, Hamilton H J, Butz C J. A foundational approach to mining item set utilities from database//Proceedings of the 4th SIAM International Conference on Data Mining. Orlando, USA, 2004: 215-221

[7] Yang S, Wang L, Bao X, et al. A framework for mining spatial high utility co-location patterns//Proceedings of the 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD'15). Zhangjiajie, China, 2015: 631-637

[8] Wang L, Jiang W, Chen H, et al. Efficiently mining high utility co-location patterns from spatial data sets with instance-specific utilities//Proceedings of the 22nd International Conference on Database System for Advanced Applications (DASFAA). Suzhou, China, 2017: 458-474

[9] Wang Xiao-Xuan, Wang Li-Zhen, Chen Hong-Mei, et al. Mining spatial high utility co-location patterns based on feature utility ratio. Chinese Journal of Computers, 2019, 42(8): 1721-1738(in Chinese)
(王晓璇，王丽珍，陈红梅等. 基于特征效用参与率的空间高效用 co-location 模式挖掘方法. 计算机学报，2019，42(8)：1721-1738)

[10] Wang X, Wang L. Incremental mining of high utility co-location form spatial databases//Proceedings of the IEEE International Conference on Big Data and Smart Computing. Jeju Island, South Korea, 2017: 215-222

[11] Li D, Wang S, Li D. Spatial Data Mining: Theory and Application. Berlin, Germany: Springer, 2015

[12] Shekhar S, Huang Y. Discovering spatial co-location patterns: A summary of results//Proceedings of the International Symposium on Spatial and Temporal Databases. Redondo Beach, USA, 2001: 236-256

[13] Garaeva A, Makhmutova F, Anikin I, et al. A framework for co-location patterns mining in big spatial data//Proceedings of the International Conference on Soft Computing and Measurements (SCM 2017). St. Petersburg, Russia, 2017: 477-480

[14] Akbari M, Samadzadegen F, Weibel R. A generic regional spatio-temporal co-occurrence pattern mining model: A case study for air pollution. Journal of Geographical Systems, 2015, 17(3): 249-274

[15] Bao X, Wang L. A clique-based approach for co-location pattern mining. Information Sciences, 2019, 490: 244-264

[16] Wang L, Bao X, Zhou L, et al. Mining maximal sub-prevalent co-location patterns. World Wide Web, 2019, 22(5): 1971-1997

[17] Chan H K, Long C, Yan D, et al. Fraction-score: A new support measure for co-location pattern mining//Proceedings of the 35th International Conference on Data Engineering (IEEE ICDE 2019). Macao, China, 2019: 1514-1525

[18] Yao H, Hamilton H J, Geng L. A unified framework for utility-based measures for mining itemsets//Proceedings of the ACM SIGKDD 2nd Workshop Utility-Based Data Mining. Philadelphia, USA, 2006: 28-37

[19] Ahmed C F, Tanbeer S K, Jeong B, et al. Efficient tree structures for high utility pattern mining in incremental databases. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(12): 1708-1721

[20] Yin J, Zheng Z, Cao L. USpan: An efficient algorithm for mining high utility sequential patterns//Proceedings of the KDD 2012. Beijing, China, 2012: 606-668

[21] Zeng X, Li Z, Wang J, et al. High utility co-location patterns mining from spatial dataset with time interval//Proceedings of the 4th International Conference on Image, Vision and Computing(ICIVC 2019). Xiamen, China, 2019: 628-636

[22] Zhong R, Li G, Tan K-L, et al. G-tree: An efficient and scalable index for spatial search on road networks. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(8): 2175-2189

[23] Li Z, Chen L Wang Y. G*-Tree: An efficient spatial index on road networks//Proceedings of the 35th International Conference on Data Engineering (ICDE). Macao, China, 2019: 268-279

[24] Wan Y, Zhou C. QuCOM: *K* nearest features neighborhood based qualitative spatial co-location patterns mining algorithm //Proceedings of the IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services. Fuzhou, China, 2011: 54-59

[25] Agrawal R, Ram B. A modified *K*-nearest neighbor algorithm to handle uncertain data//Proceedings of the 5th International Conference on IT Convergence and Security(ICITCS). Kuala Lumpur, Malaysia, 2015: 1-4

[26] Kasemsumran P, Boonchieng E. EEG-based motor imagery classification using novel string grammar fuzzy *K*-nearest neighbor techniques with one prototype in each of classes//Proceedings of the International Conference on Artificial Intelligence in Information and Communication (ICAIIC). Fukuoka, Japan, 2020: 742-745

LUO Jin, M. S. candidate. His current research interest is spatial data mining.

WANG Li-Zhen, Ph. D, professor, Ph. D. supervisor. Her research interests include spatial data mining, interactive data mining and big data analytics.

WANG Xiao-Xuan, Ph. D. candidate. Her current research interests include spatial database and data mining.

XIAO Qing, M. S., lecturer. Her research interests include data mining and its applications.

Background

Due to the continuous development of global positioning systems and the rapid popularization of mobile devices, data based on location information has exploded. In order to find useful knowledge from massive spatial data, spatial co-location pattern mining emerges as the times require. As an important branch of spatial data mining, spatial co-location pattern mining is commonly used as a method for exploration, characterization, and summarization of spatial data. Its main purpose is to discover the set of spatial features with neighbor relationships in geographical space. In this paper, we mainly study the problem of mining high utility co-locations from the spatial databases. Spatial high utility co-location pattern mining is developed from classical spatial co-location pattern mining, which introduces the utility of spatial data to the interested measurement of spatial co-location patterns. The existing spatial high utility co-location pattern mining methods can be divided into two categories: spatial features with utility and spatial instances with utility.

It is noteworthy that most of the existing mining algorithms for spatial high utility co-locations are mainly based on a single distance threshold to establish the neighbor relationships between instances, which makes the mining algorithm more sensitive to distance threshold and poor stability. Considering the rationality and effectiveness of k-nearest neighbor (k-NN) in the construction of neighbor relationships between spatial instances, in this paper, we introduce k-NN into spatial high utility co-locations, and it can be used to measure the neighbor relationships between spatial instances more reasonably. At the same time, combined with the actual situation, we propose the concepts of core elements and core patterns under k-NN relationships. Additionally, two spatial high utility core pattern mining algorithms are designed and optimized. For uniform and non-uniform distribution of spatial data sets, the mining methods have good scalability.

This work was supported by the National Natural Science Foundation of China (61966036, 61662086), and the Project of Innovative Research Team of Yunnan Province (2018HC019).

In recent years, the authors have been committed to the research of spatial data mining, and have made some achievements. In this work, we focus on spatial high utility co-location mining, and propose the related concepts and the efficient algorithms to mine spatial high utility co-locations.