基于 Voronoi 图的空间 Co-Location 核模式挖掘

邹目权[1),2)] 王丽珍[1)] 吴萍萍[1)] 杨培忠[1)]

[1)](云南大学计算机科学与工程系 昆明 650500)

[2)](昆明学院信息工程学院 昆明 650214)

摘 要 飞速发展的物联网技术不断催生海量带有时间和空间属性的数据集.这些数据集掀起了以空间 co-location 模式挖掘为代表的空间数据挖掘研究的高潮.传统空间 co-location 模式挖掘研究主要发现空间中频繁并置出现的特征的子集.特征在模式内部是无序的,特征之间的地位是平等的.例如,co-location 模式{看守所,刑警中队,武警中队}表示看守所附近往往存在刑警中队和武警中队,反之亦然.然而,由于空间分布密度差异显著存在,现实中存在特征地位不平等的模式,这些模式中的某些特征(核特征)附近频繁地出现其它特征(非核特征)的实例,而这些非核特征附近不一定频繁地出现核特征的实例.例如,某些肿瘤疾病与某些污染源的关系.在传统模型中,用户为了发现感兴趣的模式不得不将频繁性阈值设置得很低,以至于忽略了模式中特征的主从关系.本文聚焦于前述现象,研究在空间数据集中挖掘核特征与非核特征组成的有趣模式.首先,基于核邻居定义空间 co-location 核频繁模式(简称核模式)的概念.核邻居与最近邻息息相关,它不仅遵从地理学第一定律而且能排除无关实例的干扰.其次,提出核模式的有趣性度量理论,分析核模式具有的性质,如基于核参与率反单调性的先验原理等.再次,提出基于 Voronoi 图的核邻居计算思想,避免了传统 co-location 模式挖掘中为计算邻近关系需要用户预先设定距离阈值等问题.同时,扩展传统的对称的空间邻近关系到不对称的核邻居关系,使其与特征的不平等地位相适应.此外,针对点、线、面等不同几何形状的空间实例,提出基于四包理论的经典 Voronoi 图的扩展方法.最后,在合成数据与真实数据上对比验证了 Core Pattern Mining(CPM)算法的效果与效率.实验高效地发现了有别于经典 co-location 模式的有趣模式,它们具有可理解性.

关键词 空间数据挖掘;co-location 核模式;核邻居;Voronoi 图;核参与度

中图法分类号 TP311 **DOI 号** 10.11897/SP.J.1016.2022.01908

Mining Co-Location Core Patterns in Spatial Data Sets Based on the Voronoi Diagram

ZOU Mu-Quan[1),2)] WANG Li-Zhen[1)] WU Ping-Ping[1)] YANG Pei-Zhong[1)]

[1)](*Department of Computer Science and Engineering, Yunnan University, Kunming* 650500)

[2)](*School of Information Engineering, Kunming University, Kunming* 650214)

Abstract With the development of the Internet of Things (IoT), massive spatio-temporal data sets are collected. It has set off the climax of spatial data mining research that is represented by spatial co-location pattern mining. The traditional spatial co-location pattern mining is mainly used to find subsets of spatial feature sets from given spatial data sets. Spatial features in the subsets are frequently co-occurrence in the geographic space. For example, the co-location pattern {detention center, criminal police squadron, armed police squadron} reveals that there are always detention centers and criminal police squadrons near armed police squadrons and vice versa. That is to say, spatial features (e.g., the detention center and criminal police squadron) are generally

收稿日期:2020-02-07;在线发布日期:2022-02-11.本课题得到国家自然科学基金(61966036,61662086,62066023)、云南省科学创新团队项目(2018HC019)、昆明学院科学研究项目(XJZZ1706)资助.**邹目权**,博士研究生,高级工程师,中国计算机学会(CCF)会员,主要研究方向为空间数据挖掘、机器学习.E-mail:ynu_zmq@foxmail.com.**王丽珍**(通信作者),博士,教授,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.E-mail:lzhwang@ynu.edu.cn.**吴萍萍**,博士研究生,讲师,主要研究方向为空间数据挖掘.**杨培忠**,博士,主要研究方向为空间数据挖掘.

thought to be disordered and equal in each pattern in default. However, there are interesting patterns whose spatial features have unequal status because distribution densities of different spatial features tend to be different. In detail, some spatial features, which are called core features in this paper, are surrounded by the others, but the others may have no such properties. The distribution relationship between cancer cases and pollution sources is an example. To find this kind of patterns, which is called the co-location core pattern in this paper, users have to set low prevalence thresholds with neglection of those differences in the traditional co-location pattern mining. Interestingly, co-location core patterns reveal underlying subordinative compounds. Thus, this paper focuses on the unequal status of spatial features to propose theorems and methods to discover core features and their surrounding features in spatial data sets. Firstly, the prevalent co-location core pattern is defined on core neighbors. Core neighbors are strongly related to nearest neighbors. Not only do they obey the first law of geography well but they also eliminate noise from instances of unrelated features. Secondly, The prevalence measure theory of the core pattern is proposed. The prevalence of a co-location core pattern depends on co-occurrence ratios of core features but not on ones of surrounding features. Some properties of co-location core patterns are analyzed. For example, the Apriori property is based on the anti-monotonicity of co-occurrence ratios of core features. They are helpful to avoid exponential validation of candidate co-location core patterns. Thirdly, an idea for calculating core neighbors of all spatial instances based on the Voronoi diagram is presented, which can avoid setting a reasonable distance threshold for the spatial neighborhood relationship generation in the traditional co-location pattern mining. In addition, the classical spatial neighborhood relationship satisfies symmetry, but the core neighbor does not. It confirms the unequal status of spatial features. Furthermore, we propose a novel method to generate a Voronoi diagram where generators may be different geometries such as point, line, and polygon. It is an extension method of the classical Voronoi diagram based on concave envelope theory. Finally, the effectiveness and efficiency of the core pattern mining algorithm are analyzed by extensive comparison experiments both on synthetic and real-world spatial data sets. Some interesting co-location patterns which are different from classical co-location patterns are discovered. They are understandable.

Keywords　spatial data mining; co-location core pattern; core neighbor; Voronoi diagram; core participation index

1　引　言

随着基于位置服务(Location Based Service)的兴起,时空数据上的模式发现成为数据挖掘学者关注的重要方向.在时空模式中,诸如并置(Co-location)、共现(Co-occurrence)、交互(Interactions)、强相关(Correlations)、级联(Cascading)、序列(Sequential)或因果(Cause and effect relationship)等都与空间和时间两个维度相关.基本的共识是在时间维度上重叠或邻近的模式只有在空间维度上邻近才是有意义的[1].因此,模式是否在空间维度上频繁地并置出现是时空数据挖掘的基础,最具代表性的是 co-location 模式挖掘.近年来,空间 co-location 模式挖掘技术在诸如医学影像处理、生态环境保护、疾病疫情控制等领域得到了广泛应用[2-3].

空间co-location模式(简称 co-location 模式)挖掘旨在发现那些在空间中频繁并置出现的空间特征(简称特征,或称空间对象)的集合.例如,co-location 模式{兰类植被,湿润半湿润常绿阔叶林}表示兰类植被与湿润半湿润常绿阔叶林频繁地并置出现[4],印证了植物学家的相关知识,对珍稀兰类植被保护有着积极作用.

然而,传统 co-location 模式挖掘关注的是那些在空间中频繁并置出现的具有平等地位的特征的集合,即模式中的任意特征都频繁地并置出现在其它

特征附近，未考虑这样的情形：某些特征（核特征）附近频繁地出现另一些特征（非核特征）的实例，而这些非核特征附近不一定频繁地出现核特征的实例.

达尔文的《物种起源》灵感起源于伦敦附近一个叫作唐恩的小镇. 达尔文发现在小镇上看似不相关的猫、田鼠、土蜂、三叶草间存在较强的相关性，即猫多了会消灭大量田鼠，使土蜂种群免遭田鼠危害而大量繁衍，土蜂种群帮助三叶草授粉使三叶草数量猛增（土蜂是三叶草授粉的理想媒介，包括其它蜜蜂在内的昆虫皆不是）[5]. 探究这一现象发现三叶草（核特征）附近频繁地出现猫、田鼠、土蜂等；而猫、田鼠、土蜂等附近不一定出现三叶草，因为猫、田鼠、土蜂等在三叶草以外的区域分布更为广泛. 传统 co-location 模式难以表达这一现象.

在城市规划中，也存在类似的现象. 例如：三甲医院附近频繁地出现公交站台、花店、药店、旅馆、餐馆等；公交站台、花店、药店、旅馆、餐馆等附近不一定出现三甲医院；同时，三甲医院附近的公交站台、花店、药店、旅馆和餐馆并不一定彼此邻近，因为它们可能因为开在三甲医院的不同出入口而不在同一条街道. 传统 co-location 模式亦难以表达这一现象.

进一步分析，在达尔文的猫与三叶草生态系统中，猫和田鼠的活动半径是不一样的. 如果用传统 co-location 模式挖掘中的距离阈值方法判定邻近，那么某只猫与某株三叶草是否邻近依赖的距离阈值，应当与某只田鼠和某株三叶草是否邻近的距离阈值不同，即邻近关系应当体现出不同特征的空间分布密度差异. 此外，由于土蜂种群领地意识的作用，不同种群的土蜂是否与某些三叶草邻近的距离阈值也不同，即邻近关系应当体现同一特征的不同实例的空间分布密度差异. 相似地，以市政规划中的三甲医院附近的市政配套设施为例：公交站台辐射半径基本不超过 0.5 km，但是旅馆的辐射半径相比可能就要远得多；此外，市中心的公交站台辐射半径与郊区的公交站台的辐射半径也应存在差异.

传统空间邻近关系的计算是挖掘空间模式的基础性工作，主要有基于距离阈值的方法和无距离阈值的方法. 这些方法各有所长，但是两种方法（本文将在第 2 节相关工作中详细分析）都很难同时兼顾不同特征以及相同特征的不同实例的空间分布密度差异. 此外，值得注意的是，即使存在合理的距离阈值等参数，如何确定这些合理值同样面临着巨大的挑战.

本文提出基于 Voronoi 图[6]的核邻居对空间进行划分，能同时体现不同特征及其实例间的空间分布密度差异. 在对空间进行划分时，Voronoi 图生成元可能是点、线、多边形等不同几何形状的实例（对象），于是，我们提出基于不同几何形状的凹包上的顶点作为生成元的一种 Voronoi 多边形生成方法. 借助核邻居将 co-location 模式挖掘中经典的空间邻近关系由必须满足对称性扩展为不一定满足. 以核邻居概念为基础，提出核模式用以表示核特征附近频繁地出现非核特征的实例的情形，丰富了传统 co-location 模式挖掘的内涵，增强了现实应用性.

本文第 2 节是相关工作；第 3 节描述总体解决方案；第 4 节提出基于 Voronoi 图的核邻居及其求解方法；第 5 节形式化地描述核模式及挖掘算法；第 6 节是实验分析；第 7 节总结展望.

2 相关工作

随着大数据采集、处理技术的不断发展，空间 co-location 模式挖掘的研究愈加丰富多彩[7]. 下面针对 co-location 模式挖掘的两个主要工作——生成空间邻近关系和基于空间邻近关系生成空间 co-location 模式，讨论与论文研究相关的工作.

2.1 Co-Location 模式

传统 co-location 模式种类繁多，如频繁模式[8-9]、亚频繁模式[10]、稀有模式[4,11]、主导模式[12]、负模式[13]和压缩模式[14]等. 其中：稀有模式、主导模式、负模式和压缩模式在定义上虽然存在差异，但其实还是频繁模式的变种，亚频繁模式是最近提出来的一种新的模式.

传统 co-location 模式内部的特征地位是平等的，即模式内部的任意特征都频繁地出现在其它特征的附近. 这类模式通常以团实例为行实例. 所谓行实例是指匹配 co-location 模式的实例组成的集合，所有行实例的集合称为表实例；团实例是指其中的任意两个实例之间是邻近的[8]. 设图 1 中边表示它的两个端点代表的实例是邻近的，那么$\{A2,C2,D2\}$就是 co-location 模式$\{A,C,D\}$的一个行实例. 若所

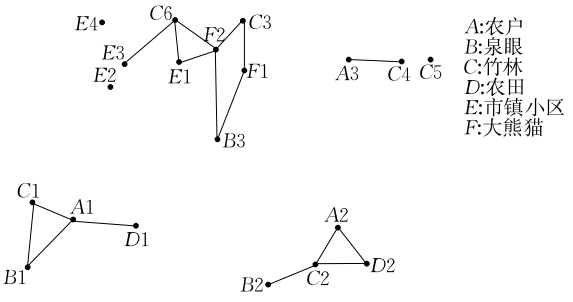


图 1 经典空间邻近关系示例

有的行实例组成的表实例中，特征的参与率（特征的实例在表实例中不重复出现的个数与它的实例数的比值）的最小值（称为参与度）大于等于用户给定的最小参与度阈值（一个 0 至 1 之间的百分数），那么就称相应的 co-location 模式为频繁模式[8]. 例如，图 1 中 co-location 模式{A,C,D}的表实例为{{$A2$,$C2$,$D2$}}，因此它的参与度为 MIN(1/3,1/6,1/3)=1/6，设最小参与度阈值为 2/3，那么此时它代表的 co-location 模式{农户，农田，竹林}不是一个频繁模式，其中 MIN 函数表示求最小值.

稀有模式、主导模式在一定程度上考虑了模式内部的特征间的不平等地位.

不同特征的实例数可能是不同的，通常将实例数极少的特征称为稀有特征. 当一个 co-location 模式中含有稀有特征时会造成它们之间地位的不平等，若使用传统 co-location 模式定义难以发现相应的模式. 为了消除这种不平等，稀有模式使用最大参与率[11]或最小加权参与率[4]代替经典模式的参与度. 然而，稀有模式的行实例与经典模式一样是团实例，未考虑核特征附近频繁地出现它们的非核特征的实例，而非核特征附近不一定频繁地出现核特征的实例的情形. 例如稀有模式{三甲医院，花店，药店}虽然可以被发现，却与现实中绝大多数的花店与药店附近不存在三甲医院相悖. 相似地，例如，图 1 中稀有特征大熊猫附近频繁地分布着竹林和泉眼，但稀有模式{大熊猫，竹林，泉眼}的表实例是空集，它的参与度为 0. 然而，观察三者之间的空间分布可以发现，虽然大熊猫附近的竹林和泉眼不一定彼此邻近，但它们确实都同时出现在了大熊猫附近，这与大熊猫以竹子为食，每天都需要饮水的生活习性是一致的.

频繁模式内部特征的地位是平等的，但在此大前提下，文献[12]认为特征间仍存在主从地位（相对不平等），即某些特征（非主导特征）的出现是由另一些特征（主导特征）主导的. 例如，图 1 中，最小参与度阈值为 2/3 时，co-location 模式{农户，农田}是一个频繁模式，根据特征重复率、特征影响度、特征关键度和模式关键度等定义可以得到农田主导了农户的出现. 然而，主导特征与非主导特征的地位之间的不平等是非常有限的，即它们首先是频繁地彼此相关，对地位相差较大的情形力不从心. 例如三甲医院虽然主导了它附近的花店和药店的出现，但由于三甲医院、花店、药店三者组成的候选模式难以确认为频繁模式，要通过主导模式的相关定义发现它们之间的主从关系无从谈起.

与频繁模式要求行实例为团实例不同，亚频繁模式[10]使用星型实例代替了团实例作为行实例，但是它仍然要求模式内部的每个特征附近都频繁地出现其它特征的实例，即模式内部的特征的地位仍然是平等的.

综上，已有的 co-location 模式挖掘方法主要针对模式内部的特征地位平等的情形. 然而，此类情形在诸如污染源对肿瘤疾病的影响研究、珍稀植被保护、城市规划等领域具有显著意义.

2.2　邻近关系

传统的 co-location 模式挖掘虽然验证模式的频繁性是其最主要的步骤，但空间邻近关系的计算才是其基础. 在不同的空间邻近关系下，即使频繁性阈值不变，挖掘得到的模式可能呈现出截然不同的结果. 因此，判定实例间是否邻近，同样受到学者们的重视. 主要的方法大致可以分为基于距离阈值的[15-16]和无距离阈值的方法[17-23].

与传统 co-location 模式挖掘相似，核模式的挖掘也需要确定什么样的情形可以视为核特征的实例（核邻居）附近出现（邻近）非核特征的实例：是核特征的某个半径范围内（基于距离阈值的方法）还是其它情形（无距离阈值方法）.

2.2.1　基于距离阈值的方法

基于距离阈值的方法主要分为静态距离阈值[15]和动态距离阈值[16]的方法，其中静态距离阈值的方法是传统 co-location 模式挖掘最常用的方法.

静态距离阈值方法操作简便，用户接受度高，但是要获得理想的效果，需要用户结合经验以及数据分布特点寻找合适的阈值. 然而如引言所述，要获得一个“好”的距离阈值是极为困难的，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空间分布密度. 相似地，若核模式挖掘使用静态距离阈值的方法，相当于认为核邻居的静态距离阈值为半径的范围内的实例都与它邻近，看似合理，但是却忽略了不同核特征的空间分布密度差异以及相同核特征的不同实例的空间分布密度差异. 如图 1 是距离阈值为 600 m 时的邻近关系图，为了发现频繁模式{农户，农田，竹林}至少要将距离阈值设置为 1000 m，然而此时 co-location 模式{大熊猫，市镇小区}变成一个频繁模式，用户显然对它不感兴趣，因为大熊猫和市镇小区在空间上应当是不相关甚至是负相关的.

动态的距离阈值方法能在一定程度上克服静态

距离阈值方法的一些不足，但需要用户做出更多的选择. 例如：文献[16]无需预先输入距离阈值、最小参与度阈值等，但是在挖掘过程中让用户不断地根据迭代生成的参考信息进行选择，过程繁琐且严重依赖用户的主观经验. 动态的距离阈值方法如果平移到核模式挖掘中，理论上存在一定的合理性，能充分地考虑到核特征以及它的实例的空间分布密度差异，但是由于相似的原因，用户体验不佳.

2.2.2 无距离阈值的方法

已有的无距离阈值方法比较有代表性的有基于最近邻、聚类、Delaunay 三角剖分等思想的方法.

基于最近邻思想的邻近关系生成较为直观，也最能直接表达 co-location 模式挖掘原理——地理学第一定律“任何事物都是与其它事物相关的，只不过相近的事物关联更紧密”[24-25]. 文献[17]使用 KNNG(k Nearest Neighbor Graph)代替距离阈值，能体现空间分布密度差异，但未区分实例所属特征，抗干扰性差. 文献[19]提出了动态邻近约束，在生成邻近关系过程中，让用户根据当时的约束条件设置缓冲区的大小，从而可以让用户自行决定邻近实例的数量，在一定程度上相对客观地反映实例在不同区域的空间分布密度，但是仍然严重依赖用户的主观经验，且与动态距离阈值类似，用户体验不佳.

聚类方法因为直观易理解成为时空模式挖掘常用方法之一，其中基于密度的聚类(DBSCAN)方法，因能有效地反映出实例的空间分布密度而使用最广[20]. 但是在邻近关系处理上，此种方法认为簇内的任意实例间都彼此可达(邻近)，簇间的实例不可达. 事实上，这混淆了实例至实例的距离和实例至簇中心的距离，显然，相邻簇的边缘上较邻近的实例间的距离可能远远小于簇内分属两个不同边缘位置的实例间的距离. 此外，诸如 DBSCAN、k-means 等聚类方法，通常都需要用户设定参数，而合理的参数设定一直是聚类方法需要突破的难题.

早期，也出现过基于事务思想生成邻近关系的方法[21]. 这种方法将 co-location 模式挖掘转化为关联分析问题，易于理解，容易被用户认可，但是因为人为地割裂了事务间原本可能存在密切相关性的实例间的联系而牺牲了实例间邻近关系的完整性.

基于 Delaunay 三角剖分的思想生成邻近关系是一种新的方法. 这种方法能有效反映不同实例的空间分布特性，且不需要用户设定阈值. 然而，此种方法由于 Delaunay 三角剖分自身的特点，文献[22]的结果呈现出难以发现高阶模式的倾向，文献[23]借助合并策略重新定义了行实例解决了高阶模式挖掘问题，但得到的候选模式参与度普遍较小，且对候选模式以外的特征的实例分布异常敏感.

综上，无论是基于距离阈值的方法还是无距离阈值方法，传统的邻近关系生成方法大多数都需要用户给定参数，而相对合理的参数通常依赖于用户的主观经验. 同时，这些方法可能在核模式挖掘中难以奏效. 受相关方法的启发，本文借鉴了最近邻思想支撑地理学第一定律，使用 Delaunay 三角剖分的对偶 Voronoi 图[26-27]，对空间进行划分.

3 总体解决方案

为发现核特征附近频繁地出现非核特征的实例，而非核特征附近不一定频繁地出现核特征的实例的模式，需要解决以下几个基本问题：(1)模式的格式；(2)“附近”的界定；(3)频繁性度量.

与传统 co-location 模式使用无序的特征的集合(例如{农户，农田})表示模式内的特征频繁并置出现相似，首先给出 co-location 核频繁模式(简称核模式)的格式. 核模式形如$\{F_1,F_2\}$，其中 F_1、F_2 分别是特征集 F 的非空真子集，且 F_1、F_2 内部的特征无序，F_1 与 F_2 交集为空集，F_1 与 F_2 是有序的($\{F_1,F_2\}$ 与 $\{F_2,F_1\}$ 代表不同的核模式). 核模式 $\{F_1,F_2\}$ 表示特征集 F_1 附近频繁地出现特征集 F_2 的实例. 为了区分 F_1、F_2 中的特征，称 F_1 中的特征为 F_2 中的特征的核特征. 例如核模式{{三叶草}，{田鼠，猫}}表示三叶草附近频繁地出现田鼠和猫. 候选核模式是有望成为核模式的模式. 相关形式化定义将在第 5 节给出.

“附近”是一个相对的概念. 例如，同样是 3000 m 距离远的居民，对沃尔玛(空间分布密度小)而言是附近，而对小区便利店(空间分布密度大)却遥不可及. 与传统空间邻近关系的定义不同，结合地理学第一定律和特征的空间分布密度差异，我们提出基于 Voronoi 图的核邻居.

经典的 Voronoi 图的生成元都是点对象，然而实例可能是点对象(如花店，称为点实例)，也可能是线对象(如河流，称为线实例)、面对象(如三甲医院，称为多边形实例)，其中线实例可以视为特殊的多边形实例. 因此，我们提出了基于多边形实例的凹包上的顶点为生成元的 Voronoi 图生成方法.

与传统 co-location 模式的行实例相对应，核模式的核行实例的格式与核模式相似，例如核模式

$\{F_1,F_2\}$的核行实例形如$\{I_1,I_2\}$，其中 I_1、I_2分别是实例集 I 的非空真子集，且 I_1、I_2内部的实例无序，I_1、I_2所属特征的集合分别包含且仅包含特征集 F_1、F_2，I_1与 I_2是有序的（$\{I_1,I_2\}$和$\{I_2,I_1\}$不同）. 核模式$\{F_1,F_2\}$的核行实例$\{I_1,I_2\}$表示 I_1中的任意实例都是 I_2中的实例的核邻居.

如果将空间实例集视为图的点集，将实例与它的核邻居用弧进行连接，使核邻居为弧尾，得到的图叫做核邻居图，后文将证明核邻居图的稠密性. 那么查找核模式的核行实例就是在核邻居图上查找匹配核模式的子图. 如果使用传统的邻接矩阵存储稠密图，直接查找匹配模式的全部子图是非常困难的[28].

与 co-location 经典模式使用参与度（模式内各特征的参与率的最小值）与用户给定的最小参与度阈值比较检验频繁性相似[4]，核模式挖掘使用核参与度（模式内核特征的核参与率的最小值）与用户给定的最小核参与度阈值比较检验核模式的频繁性.

候选核模式的数量在不剪枝时是空间特征数量的指数倍，我们将证明与经典模式先验原理类似的核模式的先验原理，从而设计候选-验证的 Core Pattern Mining（CPM）算法.

4　核邻居

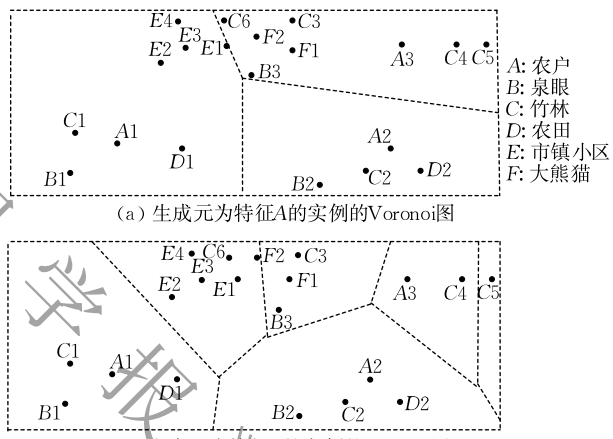
本节首先定义了基于 Voronoi 图的核邻居，其次给出了生成元为不同几何形状的实例的 Voronoi 图生成方法，再次设计了核邻居字典，最后给出了计算所有实例的核邻居的 Core Neighbor Mining（CNM）算法.

4.1　基本定义

直观地，特征的空间分布密度，可以使用所有实例到它所属特征的实例上的 k 近邻之间的距离（便于描述，称为相同特征内的最近距离，这与传统的基于 k 近邻的思想定义空间邻近关系是不同的，传统的基于 k 近邻的方法是指在所有实例当中的 k 近邻）进行刻画. 相同特征内的最近距离的均值越小，特征的空间分布密度越大；相同特征内的最近距离的方差越小，特征的空间分布越均匀. Voronoi 图是快速求解 k 近邻的重要工具[29]，因此，它在描述空间分布密度方面有着天然的优势.

定义 1.　空间实例集 I 关于特征 f_s的 Voronoi 划分 Voronoi（f_s）. 令特征 f_s的实例的集合为生成元，在空间上生成 Voronoi 图，Voronoi 多边形对空间实例集 I 的划分称为空间实例集 I 关于特征 f_s的 Voronoi 划分，记为 Voronoi（f_s）. 其中，每个 Voronoi 多边形及其合围区域称为相应生成元的 Voronoi 分块.

本文的距离使用的是欧几里得距离. 图 2 是基于图 1 得到的空间实例集关于特征的 Voronoi 划分示例，其中图 2(a)展示的是空间实例集关于空间分布相对均匀的特征 A 的 Voronoi 划分 Voronoi(A)；而图 2(b)展示的是空间实例集关于空间分布不均匀的特征 C 的 Voronoi 划分 Voronoi(C)，可以看到不同的 Voronoi 分块面积存在差别，例如 $C1$ 的 Vonoroi 分块和 $C5$ 的. 进一步地，因为特征 A 与特征 C 的空间分布密度差异较大，相应地，相同的距离对于特征 A 与特征 C 远近程度也不一样. 例如，$E2$ 到 $A1$、$C1$ 的距离相差不大，但是 $E2$ 在 $A1$ 的 Voronoi 分块中，不在 $C1$ 的 Voronoi 分块中.



(a) 生成元为特征A的实例的Voronoi图

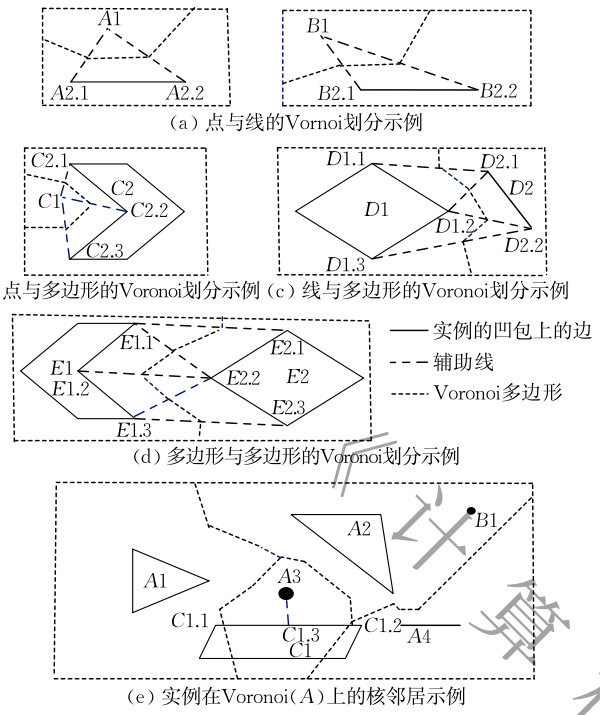
(b) 生成元为特征C的实例的Voronoi图

图 2　点实例的 Voronoi 划分示例

经典 Voronoi 图的生成元是点对象. 然而，实例不仅有点对象，还有线对象和面对象. 线对象和面对象可以使用它们的凹包描述它们的轮廓[30]. 令点实例、多边形实例到多边形实例的距离为它们之间的最近欧几里得距离.

图 3 是生成元为多边形实例时，空间实例集关于特征的 Voronoi 划分示例. 图 3(a)是邻近的两个生成元分别是点实例、线实例时对邻近区域的 Voronoi 划分示例. 它们分别展示了点到线的垂足在线上、线外的情形. 其中点到线的两个顶点构成的三角形区域使用相应三角形的等腰线进行划分，三角形外的区域分别使用点与线的各个顶点之间的中垂线进行划分. 相似地，图 3(b)是邻近的两个生成元分别为点实例、多边形实例时对邻近区域进行划分的示例；图 3(c)是邻近的两个生成元分别为线实

例、多边形实例时对邻近区域进行划分的示例；由于邻近的两个生成元分别为线实例的情形是图 3(c)的子问题，不再赘述；图 3(d)是邻近的两个生成元分别为多边形实例时对邻近区域进行划分的示例.



(a) 点与线的Vornoi划分示例

(b) 点与多边形的Voronoi划分示例 (c) 线与多边形的Voronoi划分示例

(d) 多边形与多边形的Voronoi划分示例

(e) 实例在Voronoi(A)上的核邻居示例

图 3 多边形实例的 Voronoi 划分示例

定义 2. 核邻居. 设 Voronoi(f_s)是空间实例集 I 关于特征 f_s 的 Voronoi 划分，若实例 i_u 是实例 i_v 在 Voronoi(f_s)上的 Voronoi 分块的生成元，那么称 i_u 是 i_v 在特征 f_s 上的核邻居，简称核邻居.

与作为生成元的实例相似，作为非生成元的实例可能是点对象，也可能是线对象或面对象. 令图 3(e)对应的 Voronoi 图是空间实例集关于特征 A 的 Voronoi 划分. 点实例 $B1$ 处于生成元 $A2$ 所在的 Voronoi 分块中，因此 $A2$ 是 $B1$ 在特征 A 上的核邻居. 多边形实例 $C1$ 的凹包上的顶点分别处于生成元 $A1$ 和 $A4$ 所在的 Voronoi 分块中，因此 $A1$ 和 $A4$ 都是 $C1$ 在特征 A 上的核邻居. 特别地，$A3$ 离 $C1$ 的凹包上的顶点 $C1.1$ 和 $C1.2$ 之间的边较近，检测 $A3$ 在端点分别为 $C1.1$ 与 $C1.2$ 的边上的垂足为 $C1.3$，得到 $C1.3$ 在 $A3$ 的 Voronoi 分块中，因此 $A3$ 也是 $C1$ 在特征 A 上的核邻居.

由中垂线与等腰线的几何性质可知，任意点实例或多边形实例到它所在的 Voronoi 分块对应的生成元的距离小于它到其它生成元的距离. 因此，与基于经典 Voronoi 图的空间实例集关于特征的 Voronoi 划分相似，在基于多边形的凹包上的顶点的 Voronoi 图的空间实例集关于特征的 Voronoi 划分上，任意点实例或多边形实例到它的核邻居的距离小于它到非核邻居的距离. 例如图 3(e)中，点实例 $B1$ 到 $A2$ 的距离小于它到 $A1$、$A3$、$A4$ 的距离，多边形实例 $C1$ 到 $A1$、$A3$、$A4$ 的距离小于它到 $A2$ 的距离.

显然，核邻居不一定满足对称性. 例如图 1 中，$C3$ 在特征 A 上的核邻居只有 $A3$，而 $A3$ 在特征 C 上的核邻居只有 $C4$ 无 $C3$. 这是核邻居与传统 co-location 模式挖掘中的空间邻近关系的重要区别之一. 它更符合李小文院士对空间邻近度的定义[24]，反映了特征的不平等地位.

定义 3. 核邻居图 $G(I,CNs)$. 设有向图 $G(I,CNs)$的顶点集 I 是空间实例集 I，任意实例(弧头)至它的任意核邻居(弧尾)用弧连接，所有的弧组成的集合记为 CNs，称有向图 $G(I,CNs)$是空间实例集 I 上的核邻居图，简称核邻居图.

由空间实例集关于特征的 Voronoi 划分以及核邻居的定义可知，任意实例在某个特征上的核邻居是确定的. 显然，空间实例集 I 上有且仅有一个核邻居图 $G(I,CNs)$.

设一个空间数据集的空间特征数为 m，实例数为 n，那么 $G(I,CNs)$的顶点数为 n，对应的有向完全图的弧数为 $n*(n-1)$. 由于实例在除自身以外的任意特征上至少存在 1 个核邻居，因此每个实例的出度最少为 $m-1$. 所以，$G(I,CNs)$的出度至少为 $n*(m-1)$. 一般认为，弧数 $e\geq n\log_2 n$ 时相应的图是稠密图[31]. 要使 $G(I,CNs)$为稠密图，只需 $n*(m-1)\geq n\log_2 n$，即 $m-1\geq\log_2 n$. 现实中这一情况很容易得到满足，因此，$G(I,CNs)$通常是一个稠密图. 例如，图 1 中 $n=20$，$m=6$，它对应的核邻居图是一个稠密图.

定义 4. 核邻居字典 $dict\_CNs$. 在核邻居图 $G(I,CNs)$中，以空间实例集 I 上的实例 i_v 作键，实例 i_v 在各个特征上的核邻居的有序集合作值形成的字典叫做核邻居字典，记为 $dict\_CNs$.

其中，有序是指实例按所属特征排序，相应地，使用 Python[32] 的相关数据类型进行存储. 值使用列表存储，值中的元素为每个特征上的所有核邻居(使用集合存储)，索引号为特征的序号. 例如，图 1 中有 $dict\_CNs[E1]=[\{A1,A3\},\{B3\},\{C6\},\{D1\},\{E1\},\{F2\}]$. 其中，$\{A1,A3\}$表示 $E1$ 在特征 A(排序序号为 0，对应索引号为 0)上的核邻居为 $A1$ 和 $A3$. 因此，查询实例的核邻居时间复杂度为 $O(1)$.

例如图 1 中 $dict\_CNs[A1]=[\{A1\},\{B1\},\{C1\},\{D1\},\{E2\},\{F2\}]$，$A1$ 在特征 F 上的核邻居为 $dict\_CNs[A1][5]=\{F2\}$.

4.2 Core Neighborhood Mining 算法(CNM 算法)

本小节提出 CNM 算法，求核邻居图 $G(I,CNs)$ 并输出它的核邻居字典.

4.2.1 CNM 算法

算法 1. CNM 算法.

输入：空间特征集 $F=\{f_1,f_2,\cdots,f_m\}$，实例集 $I=\{i_1,i_2,\cdots,i_n\}$ 及其空间位置坐标

输出：核邻居图 $G(I,CNs)$ 的核邻居字典 $dict\_CNs$ 表示

变量：空间实例集关于特征 f_s 的 Voronoi 划分 Voronoi(f_s)；实例的生成元(核邻居) $gener$

步骤：

1. $dict\_CNs=collections.defaultdict(list)$ //初始化字典并将值的类型定义为列表 $list$
2. $F=list(F)$
3. $F.sort()$ //特征排序
4. FOR $index,f_s$ IN $enumerate(F)$: //遍历特征并获取它的序号
5. Voronoi(f_s)=$Voronoi\_gen(f_s)$
6. FOR i IN I:
7. $gener=tree\_query(i,\text{Voronoi}(f_s))$ //查询 i 在 f_s 上的核邻居
8. $dict\_CNs[i][[index]=gener$ //将 i 在 f_s 上的核邻居有序存储
9. END FOR
10. END FOR
11. RETURN $dict\_CNs$

4.2.2 CNM 算法的时间复杂度、正确性和完备性分析

CNM 算法第 1 步初始化核邻居字典 $dict\_CNs$. 第 2～3 步对特征进行排序. 第 4～10 步是核心环节，其中：第 5 步生成空间实例集关于特征 f_s 的 Voronoi 划分；第 7 步是查询任意实例 i 在 f_s 上的核邻居；第 8 步是将核邻居有序存储到键 i 的值中. 第 11 步返回核邻居图的压缩表示——核邻居字典.

设特征集 F 上的特征数为 m，空间实例集 I 上的实例数为 n，则平均每个特征的实例数为 n/m. 为便于分析，令多边形凹包上的顶点数量为 1. CNM 算法第 5 步生成 Voronoi(f_s)的平均时间复杂度为 $O\left(\frac{n}{m}\log_2\frac{n}{m}\right)$[6]. 第 7 步借助 Voronoi 的存储结构 KDTree[6] 查询平均时间复杂度为 $O\left(\log_2\frac{n}{m}\right)$；第 8 步借助字典对核邻居进行有序存储，时间复杂度为 $O(1)$；因此，算法第 5～9 步的平均时间复杂度为 $O\left(n\log_2\frac{n}{m}\right)$. 综上，算法总的平均时间复杂度为 $O\left(mn\log_2\frac{n}{m}\right)$. 由第 8 步可知，$dict\_CNs$ 的一个键值对存储空间为 $O(m)$，因此它将空间复杂度为 $O(n^2)$ 的邻接矩阵压缩为 $O(mn)$. 由于在空间大数据集上 $n\gg m$，因此核邻居字典有效地节省了存储空间. 观察算法第 4～10 步可知，算法支持并行计算，即对 m 个 Voronoi(f_s)并行计算，此时时间复杂度可达到 $O\left(n\log_2\frac{n}{m}\right)$.

CNM 算法第 4～10 步的外循环保证了任意实例在任意特征上的核邻居都被查询，其中第 7 步保证了查询的都是核邻居，因此算法能完整到查询到任意实例的所有核邻居且核邻居字典中不存在不是核邻居的实例.

5 核模式

本节定义了核模式，用以发现核特征附近频繁地出现非核特征的实例的空间相关性模式，证明了核模式具有的基本性质，设计了 Core Pattern Mining(CPM)算法，它借鉴了 Apriori-like 算法逐阶迭代生成候选模式再验证的思想，能对候选核模式有效剪枝.

5.1 基本定义

定义 5. 候选核模式 *Candidate Core Pattern* (*CCP*). 若模式 $c=\{F_1,F_2\}$(其中 $F_1,F_2\subset F$)的第 1 个元素 F_1 和第 2 个元素 F_2 是特征集 F 的两个互不相交的非空真子集，则称模式 c 为一个候选 co-location 核频繁模式，简称候选核模式(*CCP*)，其中 F_1 称为前驱，F_2 称为后继，F_1 中的特征称为 F_2 中特征的核特征.

便于描述，$c[0]$表示候选核模式 c 的前驱，$c[1]$表示后继，则有 $c[0]=F_1$，$c[1]=F_2$. 候选核模式 c 中的特征的个数称为 c 的阶. 例如，图 1 中$\{\{A\},\{B,C\}\}$是一个 3 阶候选核模式.

定义 6. 候选核模式 c 的核行实例 $core\_row(c)$. 设 I'_1、I'_2 分别是空间实例集 I 的非空真子集，且 I'_1 中的任意实例都是 I'_2 中的任意实例的核邻居，若 I'_1、I'_2 中没有任何一个超集分别包含且仅包含 $c[0]$、$c[1]$ 中的特征，那么就称有序集合 $I'=\{I'_1,I'_2\}$ 是候选核模式 c 的一个核行实例.

其中，I'_1 称为 I'_2 的前驱实例，I'_2 称为 I'_1 的后继实

例. 候选核模式 c 的所有核行实例组成的集合称为 c 的核表实例,记为 $core\_table(c)$.

例如,图 1 中候选核模式$\{\{A\},\{B,C\}\}$的核表实例为$\{\{\{A1\},\{B1,C1\}\},\{\{A2\},\{B2,C2\}\},\{\{A3\},\{B3,C3,C4,C5,C6\}\}\}$.

定义 7. 核参与率 CPR. 设 f_s 是候选核模式 c 中的核特征,则 f_s 在候选核模式 c 的核表实例中不重复出现的个数与特征 f_s 的实例个数的比值称为特征 f_s 在候选核模式 c 上的核参与率,记为 $CPR(c,f_s)$.

设 c 为候选核模式,特征 $f_s\in c[0]$,$core\_table(c)$ 为候选核模式 c 的核表实例,$CPR(c,f_s)$ 的形式化表示:

$$CPR(c,f_s)=\frac{|\pi_{f_s}(core\_table(c))|}{|f_s|},$$

其中"π"表示关系的投影操作. 例如,图 1 中候选核模式$\{\{A\},\{B,C\}\}$的核表实例中核特征 A 出现了 $A1$、$A2$、$A3$,因此它的核参与率 $CPR(\{\{A\},\{B,C\}\},A)=3/3$.

定义 8. 候选核模式 c 的核参与度 CPI. 候选核模式 c 的核参与率的最小值称为候选核模式 c 的核参与度 CPI,记为 $CPI(c)$. 即

$$CPI(c)=\min_{f_s\in c[0]}\{CPR(c,f_s)\}.$$

例如,图 1 有 $CPI(\{\{A\},\{B,C\}\})=\min(CPR(\{\{A\},\{B,C\}\},A))=1$.

定义 9. 若 min_cpi 是用户给定的最小核参与度阈值,c 是一个候选核模式,那么当且仅当候选核模式的核参与度 $CPI(c)\geq min\_cpi$ 时,c 是一个 co-location 核频繁模式,简称核模式,记为 $c\in CPP$.

由核模式的定义可知,核模式只关注核特征的核参与率,而未考虑非核特征在核特征附近以外的实例比例. 原因在于核模式表示的是核特征的附近频繁地出现非核特征的实例. 例如若$\{\{大熊猫\},\{竹林,泉眼\}\}$是一个核模式,那么只要找到大熊猫就极可能发现它附近分布着竹林和泉眼,而不用在意这些竹林和泉眼在大熊猫附近以外的实例分布. 这也是核模式与传统 co-location 模式之间的显著区别之一.

由候选核模式的定义可知,理论上特征集 F 的任意两个不相交的非空子集组成的有序集合都可能成为候选核模式. 设特征数为 m,则所有的候选核模式数量为 $\sum_{k=1}^{m-1}\sum_{l=1}^{m-k}c_m^k*c_{m-k}^l$,可分解为 $c_m^1*(2^{m-1}-1)+c_m^2*(2^{m-2}-1)+\cdots+c_m^{m-1}*(2^1-1)$,它是指数级的. 其中,$c_m^k$ 表示在长度为 m 的元素中求 k 个元素的组合.

为有效剪枝,我们将给出核模式的基本性质和求解算法.

5.2 CPM 算法

5.2.1 核模式的基本性质

性质 1. 核参与率与核参与度的单调性. 当后继一定时,随着模式阶的增长,原有核特征的核参与率不变,核参与度单调递减;当前驱一定时,核参与率和核参与度随着模式的阶的增大而单调递减.

证明. 设有候选核模式 c,c'. 当 $c'[1]=c[1]$,$c'[0]\subset c[0]$ 时:因为核行实例中前驱实例的任意实例都是后续实例的核邻居,假设某个实例包含在候选核模式 c 的核行实例的前驱实例中,那么该实例也一定被包含在 c' 的核行实例的前驱实例中. 反之,由于任意实例在任意核特征上的核邻居是不变的,因此若某个实例包含在候选核模式 c' 的核行实例的前驱实例中,那么该实例也一定被包含在 c 的核行实例的前驱实例中. 因此,$\forall f_i\in c'[0]$:$CPR(\{c[0],c[1]\},f_i)=CPR(\{c'[0],c'[1]\},f_i)$.

又假设 $c'[0]=\{f_1,f_2,\cdots,f_k\}$,$c[0]=\{f_1,f_2,\cdots,f_k,f_{k+1}\}$,则

$$\begin{aligned}
&CPI(c)=CPI(\{c[0],c[1]\})\\
&=\min_{i=1}^{k+1}\{CPR(\{c[0],c[1]\},f_i)\}\\
&=\min\{\min_{i=1}^{k}\{CPR(\{c'[0]\cup\{f_{k+1}\},c[1]\},f_i)\},\\
&\quad CPR(\{c[0],c[1]\},f_{k+1})\}\\
&=\min\{\min_{i=1}^{k}\{CPR(\{c'[0],c[1]\},f_i)\},\\
&\quad CPR(\{c[0],c[1]\},f_{k+1})\}\\
&\leq\min_{i=1}^{k}\{CPR(\{c'[0],c[1]\},f_i)\}\\
&=\min_{i=1}^{k}\{CPR(\{c'[0],c'[1]\},f_i)\}\\
&=CPI(c').
\end{aligned}$$

当 $c'[0]=c[0]$,$c'[1]\subset c[1]$ 时:假设某个实例包含在 co-location 模式 c 的核行实例的前驱实例中,那么该实例也一定被包含在 c' 的核行实例的前驱实例中,反之则不然. 因此,$\forall f_i\in c'[0]$:

$$\begin{aligned}
CPR(\{c[0],c[1]\},f_i)&\leq CPR(\{c[0],c'[1]\},f_i)\\
&=CPR(\{c'[0],c'[1]\},f_i).
\end{aligned}$$

又设 $c'[1]=\{f_1,f_2,\cdots,f_k\}$,$c[1]=\{f_1,f_2,\cdots,f_k,f_{k+1}\}$,则

$$\begin{aligned}
&CPI(c)=CPI(\{c[0],c[1]\})\\
&=\min_{f_i=c[0]}\{CPR(\{c[0],c[1]\},f_i)\}\\
&=\min_{f_i\in c[0]}\{CPR(\{c'[0],c'[1]\cup\{f_{k+1}\}\},f_i)\}\\
&\leq\min_{f_i\in c'[0]}\{CPR(\{c'[0],c'[1]\},f_i)\}\\
&=CPI(\{c'[0],c'[1]\})=CPI(c').
\end{aligned}$$

证毕.

性质 2. 核模式的先验原理. 当且仅当一个候选核模式是核模式时,该核模式的后继不变的所有

子模式都是核模式；当一个候选核模式的前驱一定时，它的子模式中存在一个模式不是核模式时，该候选核模式不是核模式.

证明.　设有候选核模式 c,c'. 当 $c'[1]=c[1]$，$c'[0]\subset c[0]$ 时：若 c 是一个核模式，那么有 $CPI(c)=\min_{f_i\in c[0]}\{CPR(c,f_i)\}\geq min\_cpi$. 又由核参与率的单调性证明可知，

$$f_i\in c'[0]:CPR(\{c'[0],c'[1]\},f_i)=CPR(\{c[0],c'[1]\},f_i)\geq min\_cpi.$$

因此，任意的 c' 是一个核模式. 反之亦然.

当 $c'[0]=c[0]$，$c'[1]\subset c[1]$ 时：由题可知 $\exists c'\subset c:CPI(c')=\min_{f_i\in c'[0]}\{CPR(c',f_i)\}<min\_cpi\Rightarrow\exists f_i\in c'[0]:CPR(c',f_i)<min\_cpi$，由核参与率与核参与度的单调性可知，$\forall f_i\in c':CPR(c,f_i)\leq CPR(c',f_i)$. 因此，$\exists f_i\in c'[0]:CPR(c,f_i)<min\_cpi$.

证毕.

5.2.2　核模式验证

由候选核模式 c 的核表实例的定义可知，在核邻居图 $G(I,CNs)$ 上查找候选核模式 c 的核表实例 $core\_table(c)$，等价于在图 $G(I,CNs)$ 上查找匹配 c 的所有极大二部有向子图[33]. 查找所有满足条件的子图的时间复杂度往往是指数级的. 本小节将给出一个基于核邻居字典[34]的验证方法并分析复杂度.

算法 2.　C_check 算法.

输入：候选核模式 c，核邻居字典 $dict\_CNs$，最小核参与度阈值 min_cpi，特征的实例数字典 fea_count，特征的字典序 f_index，实例所属特征的字典 ins_fea

输出：BOOL 值

变量：实例在前驱上的核邻居 $core\_neibs$，特征在核表实例中出现的实例记录 cpr_core.

步骤：

```
1. core_neibs, cpr_core = collections.defaultdict(set) //初始化
2. FOR f_i IN c[1]: //遍历非核特征
3.     FOR f_j IN c[0]: //遍历核特征
4.         FOR i IN f_i: //遍历非核特征的实例
5.             core_neibs[f_i].add(dict_CNs[i][f_index[f_j]]) //添加实例 i 在 f_j 上的核邻居
6.         END FOR
7.     END FOR
8. END FOR
9. count_neibs = Counter(core_neibs.values()) //统计非核特征在前驱上的核邻居出现的次数
10. core_table = filter(lambda x: count_neibs[x] == len(c[1]), core_neibs.keys()) //筛选核表实例中的前驱实例
11. FOR core_row IN core_table: //遍历前驱实例
12.     FOR i IN core_row: //遍历核邻居
13.         cpr_core[ins_fea[i]].add(i) //按核特征记录核邻居
14.     END FOR
15. END FOR
16. CPI = MIN(len(v)/fea_count[k]
        FOR k,v IN cpr_core.items() //计算核参与度
17. IF CPI ≥ min_cpi: //验证频繁性
18.     RETURN TRUE
19. END IF
20. RETURN FALSE
```

C_check 算法第 1 步初始化相关变量；第 4～6 步求非核特征 f_i 的实例 i 在核特征 f_j 上的核邻居 $dict\_CNs[i][f\_index[f_j]]$；第 3～7 步求非核特征 f_i 在前驱 $c[0]$ 上的所有核邻居；第 2～8 步求所有非核特征在前驱 $c[0]$ 上的所有核邻居；第 9 步统计前驱 $c[0]$ 上的所有核邻居在非核特征中出现的次数；第 10 步如果前驱 $c[0]$ 上的所有核邻居在每个非核特征上都出现了 1 次，说明它是 1 个前驱实例，被筛选出来记入 $core\_table$ 中；第 11～16 步根据核参与率及核参与度的定义计算核参与度；第 17～20 步计算候选核模式的频繁性并返回结果.

C_check 算法第 5 步保证了任意添加到核行实例的前驱实例中的实例都是核邻居；第 2～8 步保证了任意非核特征在前驱上的核邻居（候选的前驱实例）都被完整且正确地记录；第 9～10 步保证了任意候选的前驱实例只有在每个非核特征上都出现了 1 次才可能被选中. 因此，被选中的候选前驱实例中的任意实例一定是任意非核特征上的某个或某几个实例的核邻居，保证了被排除的元素都不可能符合核行实例的定义. 由第 2～10 步可知，被选中的候选前驱实例一定是某个核行实例的前驱实例. 第 11～16 步保证了所有的核行实例的前驱实例都正确地参与了参与度的计算；第 17～20 步保证了候选核模式的频繁性被正确地返回. 综上，c_check 算法能正确地将候选核模式 c 的频繁性返回.

令待验证的候选核模式是 $u+v$ 阶的，其中前驱 u 阶，后继 v 阶，总实例数为 n，总特征数为 m，则平均每个特征的实例数为 n/m. c_check 算法第 5 步在 $O(1)$ 时间复杂度获取实例 i 在 f_j 上的核邻居. 第 4～6 步的平均时间复杂度为 $O(n/m)$. 第 3～7 步对 u 个核特征进行了遍历，平均时间复杂度为 $O(un/m)$. 第 2～8 步对 v 个后继进行了遍历，平均时间复杂度为 $O(uvn/m)$. 算法第 9 步统计了 v 个非核特征上

平均长度为 u 的核邻居并进行判断，平均时间复杂度为 $O(uvn/m)$. 第 11～15 步计算核参与率与核参与度，对每个前驱实例进行了遍历，由于每个前驱实例的长度为 u，平均时间复杂度为 $O(un/m)$. 第 17～19 步时间复杂度为 $O(1)$. 综上，c_check 算法的平均时间复杂度为 $O(uvn/m)$. 在空间大数据中，$u,v \ll n/m$. 因此，理论上验证任意的候选核模式是否频繁的时间复杂度是线性阶的.

5.2.3 Core Pattern Mining(CPM)算法

传统 co-location 模式挖掘有许多算法，多数以 Appriori-like 的方式进行挖掘. 由于核模式与传统 co-location 模式的格式存在显著差异，本文基于核参与率与核参与度的单调性和核模式的先验原理，提出基于逐阶递增候选-验证的 CPM 算法.

算法 3. CPM 算法.

输入：最小核参与度阈值 *min_cpi*，特征集 F

输出：核模式 *CPP*

变量：具有相同前驱的核模式的字典 *dict_head*，具有相同后继的核模式的字典 *dict_tail*，求变量 v 的元素的并集 *union*(v).

步骤：

```
1. CPP=set() //初始化为空集
2. dict_head,dict_tail,
   new_head=collections.defaultdict(set)
3. FOR f_pre IN F:
4.     FOR f_pro IN F-{f_pre}:
5.         IF c_check({{f_pre},{f_pro}}): //2 阶核模式
6.             dict_tail[{f_pro}].add({f_pre})
7.             dict_head[{f_pre}].add({f_pro})
8.             CPP.add({{f_pre},{f_pro}})
9.         END IF
10.    END FOR
11. END FOR
12. WHILE TRUE:
13.    new_head=dict_head.copy()
14.    dict_head.clear()
15.    FOR k,v IN new_head.items():
16.        Ck=ck_gen(v) //核模式先验原理
17.        IF NOT Ck:
18.            CONTINUE
19.        END IF
20.        FOR c IN Ck:
21.            IF c_check({k,c}):
22.                CPP.add({k,c})
23.                dict_head[k].add(c)
24.                dict_tail[c].add(k)
25.            END IF
26.        END FOR
27.    END FOR
28.    IF NOT dict_head:
29.        BREAK
30.    END IF
31. END WHILE
32. FOR k,v IN dict_tail.items():
33.    v=union(v)
34.    CPP.add({v,k}) //核模式先验原理
35. END FOR
36. RETURN CPP
```

CPM 算法第 3～11 步保证了对每一个可能的 2 阶候选核模式都进行了正确地验证. 其中，第 6～7 步对被验证的核模式都按照前驱和后继进行了分类记录；第 8 步保证了被验证的 2 阶核模式被完整地正确地添加到 *CPP* 中. 第 12～31 步得到了前驱为 1 阶，后继由 2 阶开始逐阶迭代增长的所有核模式. 其中，第 16～19 步利用了核模式的先验原理(前驱相同时的情形)由前驱相同的 k 阶核模式生成 $k+1$ 阶候选核模式，保证了所有被删除的候选核模式都不可能是核模式；第 20～26 保证了保留的候选核模式都被逐一使用 c_check 算法验证，其中第 22 步保证了所有核模式都被记录，第 23～24 步保证了被验证的核模式按前驱和后继分别进行了分类存储. 因此第 12～31 步保证了所有的前驱为 1 阶的核模式都被完整地正确地发现. 第 32～35 步保证了所有前驱为 1 阶的核模式在核模式先验原理(后继相同时的情形)的作用下得到的后继一定时前驱为最高阶的核模式(它代表了它所有的子模式都是核模式). 综上，CPM 算法能正确、完备地发现所有的核模式.

令空间特征数为 m，实例数为 n. 由 c_check 算法的时间复杂度 $O(uvn/m)$ 结合 $u,v \ll n/m$ 且 2 阶候选核模式的数量为 $m(m-1)$，可得 CPM 算法第 3～10 步的时间复杂度为 $O(mn)$. 算法第 12～31 步由于各阶得到的核模式数量难以确定，难以分析平均时间复杂度. 在最好情况下，生成的 2 阶核模式数量为 0 时其时间复杂度为 $O(1)$；最坏情况下，任意候选核模式都是核模式，则需要验证 $\sum_{k=2}^{m-1}\sum_{l=1}^{m-k} c_m^k * c_{m-k}^l$ 个候选核模式，时间复杂度是指数级的. 算法第 32～35 步的时间复杂度与第 12～31 步的结果有关. 因此，CPM 算法最好情况下是 $O(mn)$ 的时间复杂度，最坏情况下是指数级的.

以图 1 为例，CPM 算法的执行过程如图 4 所示.

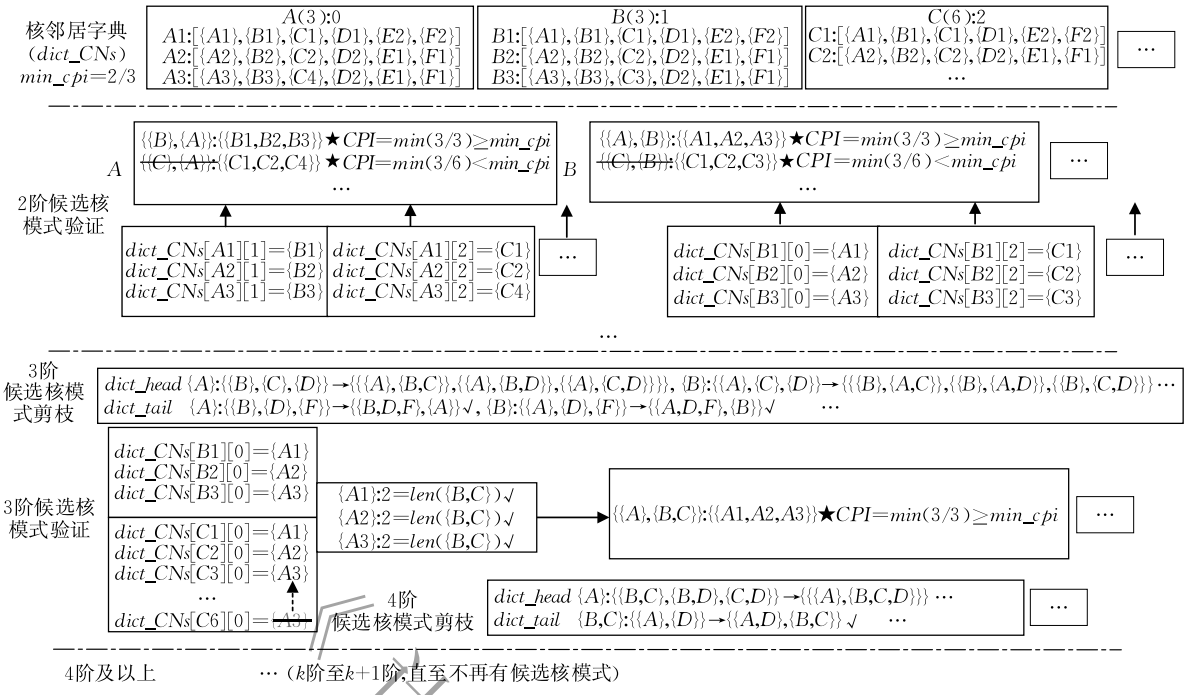


图 4　CPM 算法过程示例

6　实验分析

实验使用一台 Inter(R) Core(TM) i7-8700 CPU 3.2GHz,内存 16GB 的电脑,64 位 Windows 10 操作系统,语言 Python3.7.

实验的数据为三江并流珍稀植被数据(简称珍稀植被数据)、三江并流某区域植被数据(简称植被数据)、高黎贡山植被数据(简称高黎贡山数据)、某地癌症病例与污染企业分布数据(简称病例-污染企业数据).珍稀植被数据特征数 $m=42$,实例数 $n=336$;植被数据特征数 $m=208$,实例数 $n=3912$;高黎贡山数据特征数 $m=25$,实例数 $n=13349$;病例-污染企业数据的病例数为 5868 例,污染企业 96 家.3 个植被数据代表了 3 种不同空间分布特点的数据,其中:珍稀植被数据在空间上呈带状分布,且空间分布密度较小;植被数据在空间上呈簇状分布,空间相关性较强;高黎贡山数据在空间上呈不规则分布.

实验主要对比 CPM 算法与 Join-based 算法、DTC 算法在效果与效率上的差别.Join-based 算法[35]是传统 co-location 模式挖掘的最有代表性的算法之一.Join-based 算法是在对称的空间邻近关系上使用基于 apriori-like 的思想逐阶迭代生成候选模式再验证的方法.DTC[23]是新近发表的基于 Delaunay 三角剖分思想的无距离阈值的空间邻近关系上使用 Apriori-like 的思想逐阶迭代生成候选模式再验证的算法.值得注意的是,CPM 算法与 Join-based 算法和 DTC 算法有着本质的区别:一是 CPM 算法基于非对称的空间邻近关系,而后两者基于对称的;二是 CPM 算法的所有候选模式是空间特征集上部分有序的集合(前驱和后继分别是特征集上的组合,前驱与后继之间是排列),后两者是无序的集合(特征集上的组合).

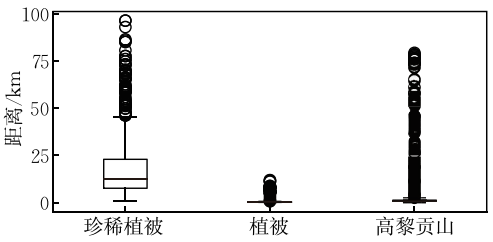
为便于描述,将三种算法的最小参与度阈值、最小核参与度阈值统称为频繁性阈值,相应地,参与度或核参与度称为频繁性值.

6.1　最近距离分析

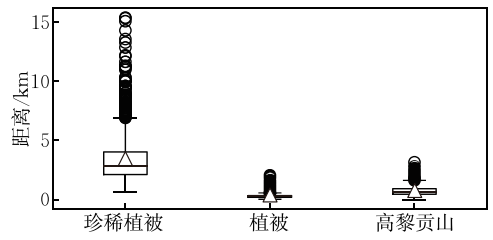
如前所述,相同特征内的 k 近邻能有效刻画特征的空间分布密度.图 5(a)是 3 个植被数据集上的相同特征内的最近距离的箱线图($k=4$).

3 个数据集上的最近距离的均值都较大,说明 3 个数据集上各个特征的空间分布密度都较小,即分布稀疏,其中珍稀植被分布最为稀疏;3 个数据集上的最近距离的方差较大,存在一定比例的异常值,说明 3 个数据集上各个特征的空间分布密度存在一定差异,其中珍稀植被数据和高黎贡山数据差异最大,植被数据相对均匀.

图 5(b)是 3 个植被数据集上各个实例到它在所有实例上的最近邻的距离的箱线图.显然地,它在各个数据集上的最近距离的均值远小于图 5(a)中的均值,这说明邻近的两个相同特征的实例之间存在一定数量的其它特征的实例.对比图 5(a)与图 5(b)



(a) 植被的k个同类最近邻的距离分布



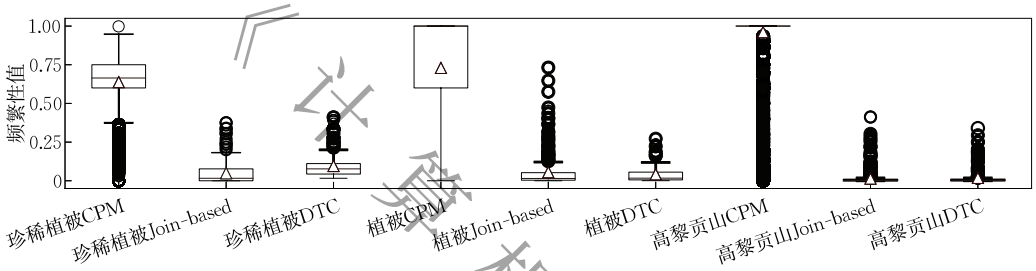
(b) 植被的k个最近邻的距离分布

图 5 实例到 k 近邻的距离分布展示($k=4$)

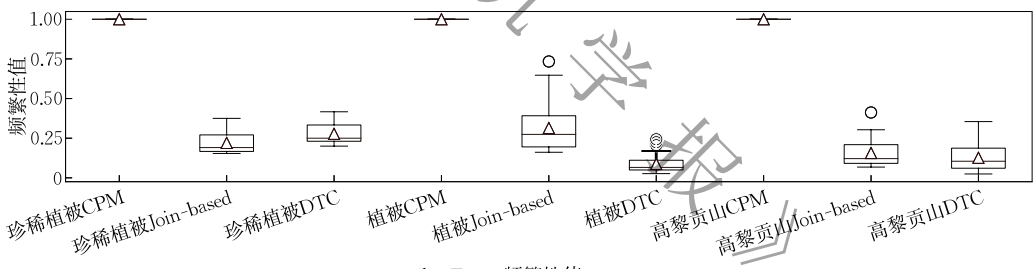
可知,珍稀植被数据和高黎贡山数据都存在较大比例的高于均值的异常值,这些异常值表明相应的特征间的空间相关性较弱.同时,在高黎贡山数据与植被数据中,均值低于中位数甚至接近 1/4 位数,说明相关数据中可能存在一定比例的空间相关性较强的模式,而珍稀植被数据相反说明空间相关性较强的模式可能较少.

6.2 频繁性分析

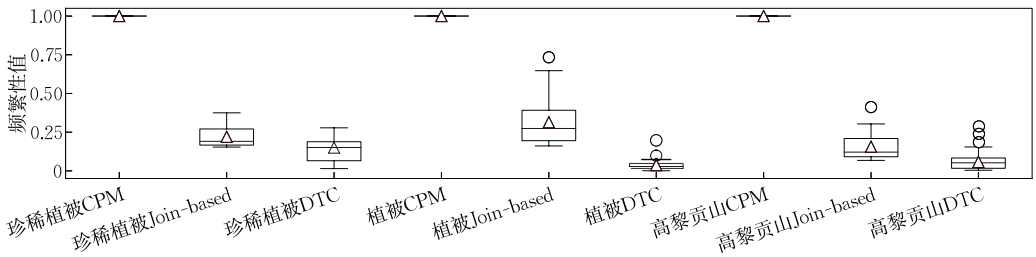
图 6(a)展示了 3 个数据集上的候选模式的频繁性值(频繁性阈值在 Join-based、DTC、CPM 上分别为 0.01、0.01、0.6,这样设置是因为三者空间邻近关系定义的不同导致它们频繁性值分布呈现出的较大差异,即 Join-based、DTC 算法对应的频繁性值往往较小,而 CPM 算法因为对应的核邻居图是稠密图而较大)的箱线图.



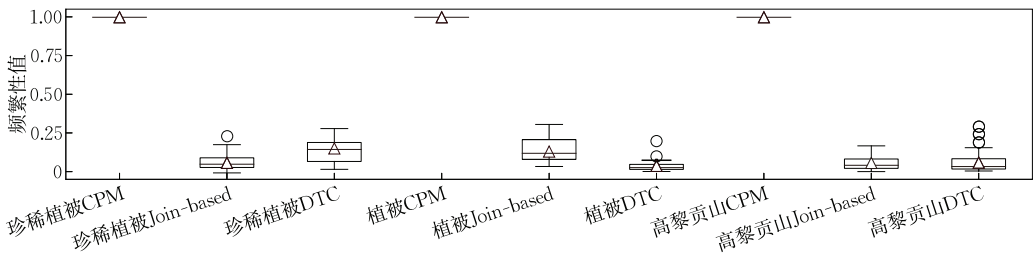
(a) 频繁性值



(b) Top30频繁性值



(c) 插入无关实例后的top30频繁性值



(d) 降低距离阈值后的top30频繁性值

图 6 模式的频繁性值及其变化箱线图

理论上，模式内部的特征间的空间相关性越高，模式的频繁性值越大. 由箱线图的特点可知，若频繁性值的箱线图中存在大量大于均值的异常值或均值大于 3/4 分位数，说明存在一定量的频繁性值特别高的模式. 那么这些频繁性值特别高的模式代表着某些特征间较强的空间相关性. 反之，若频繁性值的箱线图中存在大量小于均值的异常值或均值小于 1/4 分位数，说明存在一定量的频繁性值特别低的模式. 那么这些频繁性值特别低的模式代表着某些特征间较弱的空间相关性.

由图 6(a)可知，Join-based 算法在 3 种植被数据上都呈现出大于均值的异常值较多，均值接近 1/4 分位数，与 6.1 小节的分析一致. 因此，Join-based 在植被数据集上效果较好. 相似地，DTC 算法在植被数据上效果也较好. Join-based 算法在珍稀植被上同样存在大量大于均值的异常值，与 6.1 小节的分析结果不一致，说明它在珍稀植被上效果不理想. 相似地，CPM 算法在珍稀植被数据上效果也不太理想.

CPM 算法在珍稀植被上频繁性值相对较低，且存在大量低于均值的异常值；在另外 2 个数据集上大量频繁性值接近 1 且同样存在大量小于均值的异常值. 这与 6.1 小节的分析结论是一致的. 因此，CPM 算法的结果更能有效地反映特征间的空间相关性.

6.3　鲁棒性分析

空间 co-location 模式挖掘的结果除了特征的空间分布外，还会受到特征数、实例数、距离阈值和频繁性阈值的影响. 其中，距离阈值和频繁性阈值是影响最显著的两个因素. 图 6(b)是图 6(a)中筛选的频繁性值由高到低的 top30 个模式的频繁性值的箱线图.

若特征间存在较强的空间相关性，那么新增加一定比例的无关实例后未引起原有实例的增加或消除，理论上应当不影响原来的空间相关性. 图 6(c)是在图 6(b)的基础上增加一定比例的无关实例后，原来的 top30 个模式的频繁性值的箱线图. 对比图 6(c)和图 6(b)可知，Join-based 和 CPM 算法的频繁性值未发生改变，这说明两者对无关实例的鲁棒性优于 DTC 算法.

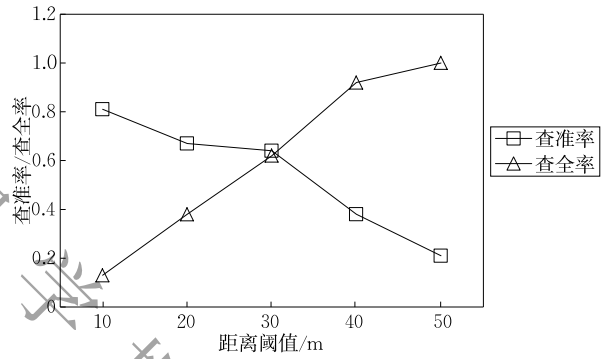
图 6(c)相较图 6(b)增加了一定比例的实例，实例的空间分布密度增大了，因此，对于 Join-based 算法应当降低距离阈值. 对比图 6(c)与图 6(d)可知，Join-based 算法得到的频繁性值再次下降了. 由对图 6(c)的分析可知，这与原有的空间相关性不变的预期相悖.

综上，CPM 算法相较 Join-based 算法和 DTC 算法对距离阈值和无关实例的鲁棒性更强，更能反映特征间的空间相关性.

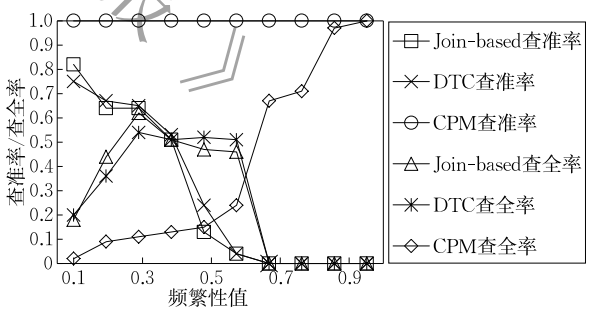
6.4　查准率与查全率分析

我们合成了含有 20 组存在明显空间相关性的模式的数据. 在这些模式中，模式内的特征与模式外的特征间彼此相对独立. 同时，在空间中随机插入了一定比例的无关实例. 理想的情况下，期望得到的频繁模式应当包含且仅包含这 20 组模式，即查准率与查全率都是 100%.

图 7(a)是频繁性阈值为 0.3 时不同距离阈值对 Join-based 算法的查准率与查全率的影响. 在频繁性阈值一定时，随着距离阈值的增大，Join-based 算法的查准率趋势是下降的，而查全率稳步上升直到 100%. DTC 算法与 CPM 算法由于不需要输入距离阈值，查准率与查全率不会发生改变，不做展示.



(a) 距离阈值对查准率与查全率的影响测试



(b) 频繁性阈值对查准率与查全率的影响测试

图 7　查准率与查全率测试

图 7(b)是 Join-based 算法的距离阈值为 30 m 时，Join-based、DTC、CPM 算法在不同频繁性阈值下查准率与查全率的变化. Join-based 和 DTC 算法的查全率随着频繁性阈值的增大趋于下降，直到为 0；Join-based 和 DTC 算法的查准率随着频繁性阈值的增大开始是上升的，但到一定程度后就下降为 0 不再变化，原因在于当频繁性阈值到达一定值后不再产生频繁模式；CPM 算法的查全率始终是 100%，原因在

于空间相关性较强的核模式的频繁性值都较高;查准率随着频繁性阈值的上升而上升,原因在于空间相关性不大的模式随着频繁性阈值的提升逐渐被排除.

对比 3 种算法的结果可知,在查准率与查全率的平衡点上 CPM 算法优于 DTC 算法和 Join-based 算法.

6.5 典型模式分析

图 8 是珍稀植被数据分布图.由于植被数据和高黎贡山数据等实例数过多,病例-污染企业数据敏感,未做展示.

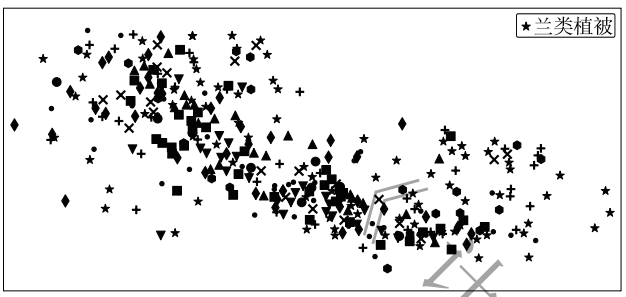


图 8 珍稀植被数据分布展示

观察图 8 可以发现,“兰类植被”(“*”形)广泛分布于整个实例采集区域,从空间分布特性的角度直观分析,“兰类植被”以外的大多数植被附近应当发现“兰类植被”[36].通过实验对比分析却发现,以 Join-based 算法为代表的传统 co-location 模式挖掘方法设置距离阈值为 2800 m,模式的频繁性阈值为 0.1 时只能发现少数 5 个特征与“兰类植被”组成的 co-location 模式;若将频繁性阈值提升至 0.2,则一个与“兰类植被”有关的 co-location 模式都不能被发现.如果使用本文提出的核模式的方法,直接将阈值设置为 0.6,可以发现有 30 个特征分别与“兰类植被”组成核模式,且“兰类植被”都作为后继,核参与度全都接近 1,较符合“兰类植被”的空间分布特性.而反过来,当“兰类植被”为候选核模式的前驱(核特征)时,其它特征做后继的特征,核参与度都特别小,这也是传统 co-location 模式挖掘方法仅能挖掘到很少带有兰类特征的 co-location 模式的原因.

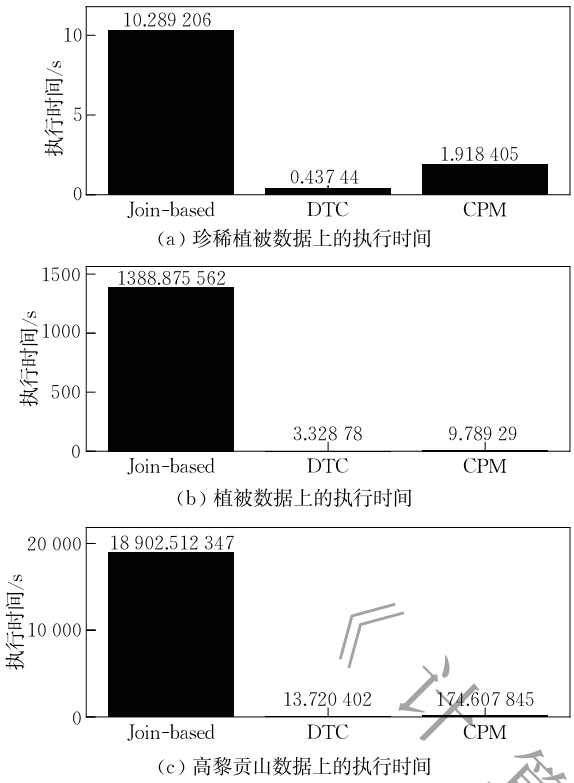
类似地,在病例-污染企业数据中,癌症 C、K、M、W 等病人常住地附近往往分布着某些种类的重点监测污染企业.如果使用经典的 co-location 模式挖掘方法,由于病例与污染企业之间的地位不平等,以及合理的距离阈值难以确定等因素,难以发现期望的模式.然而,使用基于核邻居的核模式挖掘方法,挖掘到了{{C},{甲类重点监控污染企业}}、{{K},{乙类重点监控污染企业}}、{{M},{丙类重点监控污染企业}}、{{W},{甲类重点监控污染企业}}等核模式,更高阶的如{{C,W},{甲类重点监控污染企业}}等核模式也有被发现.这在一定程度上说明某些癌症可能与某些种类的重点监控污染企业间存在一定的相关性.有趣的是,我们发现这类癌症普遍与消化道或呼吸道有关,而与乳腺、前列腺等相关性较弱.查阅相关医学资料发现,确实有些医学家认为此类癌症与污染有关[37-40].

此外,植被数据或高黎贡山数据中,核模式{{针阔混交林},{温性针叶林}}印证了“针阔混交林”与“温性针叶林”在分布上具有明显的相关性.使用传统 co-location 模式挖掘方法需要将距离阈值设置得足够大、模式的阈值设置得足够小才能发现{针阔混交林,温性针叶林}.相似地,我们观察发现“冰雪”与“温性针叶林”几乎是平行的带状分布,显然相关性非常高.相应地,我们发现了核模式{{冰雪},{温性针叶林}}和{{温性针叶林},{冰雪}}.然而,要发现传统 co-location 模式{冰雪,温性针叶林},需要将距离阈值设置得足够大且频繁性阈值足够小,而在相应参数下,大量空间相关性不大的模式(例如,{寒温山地硬叶常绿栎林,温性针叶林})也会被挖掘出来.

综合 6.1~6.4 小节分析可知,CPM 算法的挖掘结果更能有效和充分地反映空间特征之间的空间相关性.

6.6 效率对比

如前所述,CPM 算法与 DTC 算法和 Join-based 算法有着 2 个本质的区别.结合图 6(a)可知,3 种算法中 CPM 算法与 DTC 算法不需要输入距离阈值;CPM 算法因为是稠密图,所以它的候选核模式的频繁性值普遍较大,而 DTC 算法和 Join-based 算法则相对较小.因此,为了更为公平、客观地对 3 种算法的执行效率进行比较,设定参数使三者挖掘的频繁模式数量大体相当时,检测三者在 3 个植被数据集上的总执行时间进行对比,如图 9 所示.其中图 9(a)是珍稀植被数据上的执行时间,图 9(b)的是植被数据上的,图 9(c)的是高黎贡山数据上的.对比 3 种算法的执行时间可以发现 Join-based 算法执行时间最长,原因在于算法生成空间邻近关系耗时较长且候选模式是基于全连接的思想[35];CPM 算法的执行时间比 DTC 算法的长,原因在于 CPM 算法生成空间邻近关系是基于 Voronoi 图的,而 Voronoi 图的生成过程一般基于 Delaunay 三角剖分,而 DTC 算法就是基于 Delaunay 三角剖分的.



(a) 珍稀植被数据上的执行时间

(b) 植被数据上的执行时间

(c) 高黎贡山数据上的执行时间

图 9　算法在不同数据集上的总执行时间比较

对比图 9(a)、(b)、(c)中 CPM 算法在不同数据上的执行时间可知，随着实例数和特征数的增加，CPM 算法的执行时间也在增加. 这与论文中的 CNM、c_check、CPM 算法的时间复杂度分析结论是一致的.

7　总结与展望

传统 co-location 模式挖掘专注于地位平等的特征间频繁并置出现的空间相关性，未考虑地位不平等的空间相关性. 本文提出的核模式可以表示它的前驱附近频繁地出现后继的实例，并且前驱主导了后继的实例的出现，从而将 co-location 模式的应用场景进一步延伸和扩展. 同时，引入 Voronoi 图对空间进行划分，从而有效体现“附近”这一模糊的、相对的概念，并解决特征及其实例的空间分布密度差异对“附近”的影响. 为了使经典的 Voronoi 图能适应生成元为不同几何形状的实例，我们提出了基于多边形凹包上的顶点的 Voronoi 图生成方法.

实验分析说明了核模式能有效地发现空间特征间较强的空间相关性，能较好地发现核特征附近频繁地出现非核特征的模式. CPM 算法对距离阈值以及其它无关实例的鲁棒性优于以 Join-based 为代表的经典算法和最新的 DTC 算法.

虽然本文已经设计了效率相对较高的 CPM 算法，但是最坏情况下候选核模式的数量在经过剪枝后仍然是指数级的，在执行效率方面仍有不足，下一步考虑设计更加高效的算法；另外，观察发现有些核模式间相似性特别高，且核模式与它的子模式的核参与度呈现有规律的变化，如何去发现并评价这种变化从而揭示它们内在的联系也是下一步工作的内容.

参　考　文　献

[1] Moosavi S, Samavatian M H, Nandi A, et al. Short and long-term pattern discovery over large-scale geo-spatiotemporal data//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Chicago, USA, 2019: 2905-2913

[2] Akbari M, Samadzadegan 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

[3] Li J, Adilmagambetov A, Mohomed Jabbar M S, et al. On discovering co-location patterns in datasets: A case study of pollutants and child cancers. GeoInformatica, 2016, 20(4): 651-692

[4] Feng Ling, Wang Li-Zhen, Gao Shi-Jian. A new approach of mining co-location patterns in spatial datasets with rare features. Journal of Nanjing University (Natural Sciences), 2012, 48(1): 99-107(in Chinese)
(冯岭, 王丽珍, 高世健. 一种带稀有特征的空间 co-location 模式挖掘新方法. 南京大学学报(自然科学版), 2012, 48(1): 99-107)

[5] Darwin C. On the origin of species by means of natural selection. American Anthropologist, 1963, 61(1): 176-177

[6] Aurenhammer F. Voronoi diagrams—A survey of a fundamental geometric data structure. ACM Computing Surveys, 1991, 23(3): 345-405

[7] Ouyang Zhi-Ping, Wang Li-Zhen, Zhou Li-Hua. Mining spatial co-location patterns for fuzzy location of instances. Journal of Frontiers of Computer Science and Technology, 2012, 6(12): 1144-1152(in Chinese)
(欧阳志平, 王丽珍, 周丽华. 实例位置模糊的空间 co-location 模式挖掘研究. 计算机科学与探索, 2012, 6(12): 1144-1152)

[8] Ouyang Zhi-Ping, Wang Li-Zhen, Chen Hong-Mei. Mining spatial Co-location patterns for fuzzy objects. Chinese Journal of Computers, 2011, 34(10): 1947-1955(in Chinese)
(欧阳志平, 王丽珍, 陈红梅. 模糊对象的空间 Co-location 模式挖掘研究. 计算机学报, 2011, 34(10): 1947-1955)

[9] Wang L, Bao Y, Lu J, Yip J. A new join-less approach for co-location pattern mining//Proceedings of the 2008 IEEE 8th International Conference on Computer and Information Technology. Sydney, Australia, 2008: 197-202

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

[11] Huang Y, Pei J, Xiong H. Mining co-location patterns with rare events from spatial data sets. GeoInformatica, 2006, 10(3): 239-260

[12] Fang Yuan, Wang Li-Zhen, Zhou Li-Hua. Mining spatial co-location patterns with key features. Journal of Data Acquisition and Processing, 2018, 33(4): 692-703(in Chinese)
(方圆, 王丽珍, 周丽华. 含关键特征的显著 Co-location 模式挖掘研究. 数据采集与处理, 2018, 33(4): 692-703)

[13] Wu Ping-Ping, Wang Li-Zhen, Deng Shi-Kun. Research on spatial co-locations mining: A survey. Computer Science and Application, 2018, 8(3): 328-338(in Chinese)
(吴萍萍, 王丽珍, 邓世昆. 空间并置模式挖掘研究. 计算机科学与应用, 2018, 8(3): 328-338)

[14] Wang L, Bao X, Chen H, Cao L. Effective lossless condensed representation and discovery of spatial co-location patterns. Information Sciences, 2018, 436(1): 197-213

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

[16] Zhao J, Wang L, Bao X, Tan Y. Mining co-location patterns with spatial distribution characteristics//Proceedings of the 2016 International Conference on Computer, Information and Telecommunication Systems. Kunming, China, 2016: 1-5

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

[18] Qian F, He Q, Chiew K, et al. Spatial co-location pattern discovery without thresholds. Knowledge and Information Systems, 2012, 33(2): 419-445

[19] Qian F, He Q, He J. Mining spatial co-location patterns with dynamic neighborhood constraint. Information Science, 2009, 396(3): 238-253

[20] Fang Y, Wang L, Hu T. Spatial co-location pattern mining based on density peaks clustering and fuzzy theory//Proceedings of the Asia-Pacific Web and Web-Age Information Management Joint International Conference on Web and Big Data. Macau, China, 2018: 298-305

[21] Huang Y, Zhang P, Zhang C. On the relationships between clustering and spatial co-location pattern mining. International Journal on Artificial Intelligence Tools, 2008, 17(1): 55-70

[22] Sundaram V M, Paneer P. Discovering co-location patterns from spatial domain using a Delaunay approach. Procedia Engineering, 2012, 38(1): 2832-2845

[23] Tran V, Wang L. Delaunay triangulation-based spatial colocation pattern mining without distance thresholds. Statistical Analysis and Data Mining, 2020, 13(3): 282-304

[24] Li Xiao-Wen, Cao Chun-Xiang, Chang Chao-Yi. The first law of geography and spatial-temporal proximity. Chinese Journal of Nature, 2007, 29(2): 69-71(in Chinese)
(李小文, 曹春香, 常超一. 地理学第一定律与时空邻近度的提出. 自然杂志, 2007, 29(2): 69-71)

[25] Sui D Z. Tobler's first law of geography: A big idea for a small world? Annals of the Association of American Geographers, 2004, 94(2): 269-277

[26] Papadopoulou E, Lee D T. A new approach for the geodesic Voronoi diagram of points in a simple polygon and other restricted polygonal domains. Algorithmica, 1998, 20(4): 319-352

[27] Liu C H. A nearly optimal algorithm for the geodesic Voronoi diagram of points in a simple polygon. Algorithmica, 2019, 82(4): 915-937

[28] Carletti V, Foggia P, Saggese A, et al. Introducing VF3: A new algorithm for subgraph isomorphism//Proceedings of the International Workshop on Graph-Based Representations in Pattern Recognition. Anacapri, Italy, 2017: 128-139

[29] Kolahdouzan M, Shahabi C. Voronoi-based k nearest neighbor search for spatial network databases//Proceedings of the 13th International Conference on Very Large Data Bases. San Francisco, USA, 2004: 840-851

[30] Kirkpatrick D G, Seidel R. On the shape of a set of points in the plane. IEEE Transactions on Information Theory, 1983, 29(4): 551-559

[31] Dutta, S, Lauri, J. Finding a maximum clique in dense graphs via χ^2 statistics//Proceedings of the International Conference on Information and Knowledge Management. Beijing, China, 2019: 2421-2424

[32] Sanner M F. Python: A programming language for software integration and development. Journal of Molecular Graphics and Modelling, 1999, 17(1): 57-61

[33] Joyce M A, Vijaya M. Undirected binary fuzzy graphs on composition, tensor and normal products. International Journal of Research and Analytical Reviews, 2019, 6(1): 504-509

[34] Zhang P, Xing L, Yang N, et al. Redis++: A high performance in-memory database based on segmented memory management and two-level hash index//Proceedings of the 2018 IEEE International Conference on Parallel and Distributed Processing with Applications, Ubiquitous Computing and Communications, Big Data and Cloud Computing, Social Computing and Networking, Sustainable Computing and Communications. Melbourne, Australia, 2018: 840-847

[35] Yoo J S, Shekhar S. A partial join approach for mining co-location patterns: A summary of results//Proceedings of the 12th ACM International Workshop on Geographic Information Systems. Washington, USA, 2004: 241-249

[36] Yin Zhao-Lu, Liu Yu-Qing, Xiao Cui. Geographic deviation and environmental factor explanation of orchidaceae specimen record collection on the national specimen sharing platform. Science Research Information Technology and Application, 2018, 9(5): 64-71(in Chinese)

(尹朝露，刘雨晴，肖翠. 国家标本资源共享平台兰科植物标本记录采集地理偏差及其环境因子解释. 科研信息化技术与应用，2018，9(5)：64-71)

[37] Pope Iii C A，Burnett R T，Thun M J，et al. Lung cancer，cardiopulmonary mortality，and long-term exposure to fine particulate air pollution. JAMA，2002，287(9)：1132-1141

[38] Mumford J L，He X Z，Chapman R S，et al. Lung cancer and indoor air pollution in Xuan Wei，China. Science，1987，235(4785)：217-220

[39] Hill A B，Faning E L，Perry K，et al. Studies in the incidence of cancer in a factory handling inorganic compounds of arsenic. British Journal of Industrial Medicine，1948，5(1)：1-15

[40] Makris K C，Voniatis M. Brain cancer cluster investigation around a factory emitting dichloromethane. The European Journal of Public Health，2018，28(2)：338-343

ZOU Mu-Quan，Ph. D. candidate，senior engineer. His research interests include spatial data mining and machine learning.

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

WU Ping-Ping，Ph. D. candidate，lecturer. Her research interest is spatial data mining.

YANG Pei-Zhong，Ph. D. His research interest is spatial data mining.

Background

In recent years，with the continuous development of data collection and processing technologies and the rapid development of location-based services，spatial data mining has attracted high attention. As one of the important directions of spatial data mining researches，spatial co-location pattern mining has also achieved a golden period of development. Traditional spatial co-location pattern mining aims to discover the corresponding subsets of spatial features whose objects or events are located in close spatial proximity with high probability. Definitely，a variety of patterns have been defined，such as prevalent patterns，sub-prevalent patterns，patterns with rare events，and patterns with key feature. Similarly，various binary neighborhood relationships have been proposed to represent spatial proximity according to distance threshold or not.

The features are disordered and equal in each of the patterns mationed earlier. That is to say，the traditional co-location pattern mining mainly focuses on that objects or events in any one of features are in the vicinity of the others' with high probability. However，users may be interested in the pattern where objects or events in some of features are in the vicinity of the others' with high probability. Therefore，a novel pattern named core pattern has been defined in this paper.

Neighborhood relationships in the traditional co-location pattern mining satisfy symmetry，whether the distance threshold or other methods are used. However，an object is located in the vicinity of another，and the latter is not necessarily in the vicinity of the former，for example，the relationship between a panda and the bamboo forest where it lives. In this paper，we have proposed a novel directed neighborhood relationship named core neighbor based on Voronoi diagram to describe the spatial distribution density differences of features.

This work is supported by the National Natural Science Foundation of China (Nos. 61966036，61662086，62066023)，the Project of Innovative Research Team of Yunnan Province (No. 2018HC019) and the Research Project of Kunming University (No. XJZZ1706).

Our research group has been working on spatial data mining，database system，distributed and parallel computing for many years. Related works were published in good-reputation journals and conferences，such as Chinese Journal of Computer，Journal of Software，TKDE，INS and ICDE 2018 etc.