

DOI:10.13203/j.whugis20150590



文章编号:1671-8860(2017)12-1696-07

顾及属性空间分布不均的空间聚类方法 ——以城市商业中心的提取为例

朱 杰^{1,2} 孙毅中^{1,2} 陈律余^{1,2} 周 卫^{1,2} 孟耀伟^{1,2}

1 南京师范大学虚拟地理环境教育部重点实验室,江苏,南京 210023
2 江苏省地理信息资源开发与利用协同创新中心,江苏,南京 210023

摘 要:针对 Delaunay 三角网空间聚类存在的不足,提出一种顾及属性空间分布不均的空间聚类方法。首先将 Delaunay 三角网空间位置聚类作为约束条件,采用广度优先搜索方法,以局部参数“属性变化率”作为阈值识别非空间属性相似簇的聚类过程。以城市商业中心为例,验证了该方法能够更客观地识别非空间属性相似的簇,且自适应属性阈值可以满足不同聚类需求,为城市商业中心等空间实体的提取提供了一种有效方法。

关键词:空间聚类;Delaunay 三角网;属性空间分布不均;属性阈值;城市商业中心

中图法分类号:P208

文献标志码:A

空间聚类分析是空间数据挖掘与知识发现的重要内容,能够发现空间实体的空间集聚模式,揭示空间实体的分布规律。空间聚类分析进一步结合空间实体的非空间属性在空间上的分布与差异,对于解释复杂的地理现象具有重要的意义^[1]。Delaunay 三角网是空间邻近关系表达^[2,3]和空间聚类^[4-6]的一种有效工具。现有 Delaunay 三角网空间聚类方法主要是一体化法^[7,8]和空间位置上的非空间属性相似性聚类方法。这类方法将空间坐标和非空间属性数据加权融合构造距离函数,将空间位置和属性特征纳入统一的空间距离测度^[9,10],虽然能够简化变量提高效率,但由于空间位置和非空间属性之间的不可比性^[11]和属性阈值难以准确获得,影响了聚类方法的可操作性。第二类方法首先考虑实体的空间邻近性,再加入非空间属性约束,聚类结果精度较高,能够满足空间数据分布的分异特性,但属性阈值采用的是全局参数^[12,13],实质上假设了空间实体分布的均匀性,如文献[12]和文献[14]将非空间属性阈值设为最邻近实体非空间属性差异的平均值,难以满足非空间属性分布不均匀的地理现象(如城市商业中心),有可能将局部相似的对象划入到噪声簇

或者其他聚类簇中,导致聚类精度降低。针对上述问题,本文提出了一种顾及属性空间分布不均的空间聚类方法。

1 考虑属性空间分布不均的空间聚类方法

本文提出的空间聚类方法首先进行空间位置聚类,再在此基础上顾及非空间属性聚类。

1.1 空间位置聚类

采用文献[15-16]结合的聚类策略,分别对 Delaunay 三角网施加整体约束和局部约束,具体表达如下。

定义 1 给定一个空间数据集 D , 包含对象生成的 Delaunay 三角网表示为 D_T 。针对任一对象 P , 与其直接相连的对象表示为 $D_N(P)$, $L_{\text{mean}}(P)$ 是 P 与 $D_N(P)$ 相连的边长均值, 表示为:

$$L_{\text{mean}}(P) = \frac{1}{d(P)} \sum_{i=1}^{d(P)} |e_i| \quad (1)$$

式中, $d(P)$ 表示对象 P 直接邻近对象的个数; $|e_i|$ 表示 P 与 $D_N(P)$ 相连的边长。

定义 2 $L_{\text{SD}}(P)$ 是对象 P 与 $D_N(P)$ 相连边长的标准差, 表示为:

收稿日期:2016-09-14

项目资助:国家自然科学基金(41671392,41371374,41301405);公安部科技强警基础工作专项项目(2015GABJC39)。

第一作者:朱杰,博士生,主要从事空间数据表达与挖掘研究。Chu_je@163.com

通讯作者:孙毅中,博士,教授。sunyizhong_cz@163.com

$$L_{SD}(P) = \sqrt{\frac{\sum_{i=1}^{d(P)} (L_{mean}(P) - |e_i|)^2}{d(P)}} \quad (2)$$

$L_{mean}(P)$ 和 $L_{SD}(P)$ 两个指标都只反映了点的局部变化,无法满足不同密度的空间簇。

定义 3 全局性判断 G_{SD} 是 D_T 内所有对象 $L_{SD}(P)$ 的平均值,表示为:

$$G_{SD} = \frac{1}{N} \sum_{i=1}^N L_{SD}(P) \quad (3)$$

式中, N 表示空间数据集 D 所有对象的个数。

定义 4 整体边长约束。当 $|e_i|$ 大于 $L_{mean}(P)$ 加 G_{SD} 时, e_i 被视为长边,表示为:

$$L_{Edge}(P) = \{e_i \mid |e_i| > L_{mean}(P) + G_{SD}\} \quad (4)$$

如果网内任一对象 P 与 $D_N(P)$ 相连的边长长度大于或等于长边长度,则将其删除。

定义 5 局部边长约束。在完成整体边长约束的基础上, Delaunay 三角网内依然存在若干不一致的边,针对对象 P 进一步施加局部边长约束准则:

$$S = (F(P) \mid F(P) = \frac{L_{SD}(P)}{L_{mean}(P)}, F(P) \leq \gamma) \quad (5)$$

式中, $F(P)$ 表示对象 P 邻域内的紧凑程度, $F(P)$ 越小该邻域越紧凑,该参数能够探测到异常的“短边”和“长边”; γ 表示紧凑度阈值,采用文献 [15] 启发式方法确定 γ , 连接所有 $F(P) \leq \gamma$ 的点及其邻域。

1.2 非空间属性聚类

通常情况下属性空间分布是非均匀的,为适应复杂的属性空间分布情况,应要顾及非空间属性相似性的局部变化信息。为此,给出以下定义。

定义 6 属性密度。空间位置聚类后获得 Sub_D_T 空间邻近图,针对任一对象 P ,与其直接相连的全部对象包括自身表示为 $Sub_D_N(P)$,空间对象 P 的属性密度 $\rho(P) = \frac{l_{sd}/l_{mean}}{N}$, 其中, l_{mean}

$$(P) = \frac{\sum_{i=1}^K Sub_D_N(P)_{i.a}}{K}, K \text{ 表示 } Sub_D_N$$

(P) 的个数, $Sub_D_N(P)_{i.a}$ 表示 $Sub_D_N(P)$ 第 i 个对象的非空间属性值; $l_{sd}(P) =$

$\sqrt{\sum_{i=1}^K (Sub_D_N(P)_{i.a} - l_{mean}(P))^2 / K}$ 。 l_{sd}/l_{mean} 是一个相对参数,能够比较量纲不同或者平均值差别较大的两组数据的离散度,其值越小说明该对象属性离散度越小; N 表示 $D_N(P)$ 的个数,可以用来表示对象的邻域密度大小; $\rho(P)$ 既考虑了对象直接邻域内的属性离散度,又顾及了

对象的邻域密度大小,离散度小且密度大的对象属性密度越大,属性空间分布越紧凑。

定义 7 聚类中心点。在 Sub_D_T 邻近图中, C_P 表示聚类中心点,表示为:

$$C_P(P) = \min(\rho(P)) \quad (6)$$

计算所有对象 $\rho(P)$ 值,按照升序排列,取最小值作为聚类中心点。聚类中心不是固定的,当一个聚类簇形成时,在所有未被聚类的对象中,按照上述定义重新选择中心点。

定义 8 直接邻居。给定一个对象 P ,存在一个对象 $Q \in D_N(P)$,则 Q 是 P 的直接邻居。

定义 9 间接邻居。如果存在一个对象链 $P_1, P_2, \dots, P_{n-1}, P_n$, 满足 P_n 仅属于 $D_N(P_{n-1})$, P_k 属于 $\{D_N(P_{k-1}) \cap D_N(P_{k+1}), 1 < k < n\}$, P_2 仅属于 $D_N(P_1)$, 那么 P_3, P_4, \dots, P_n 是 P_1 的间接邻居。

定义 10 属性变化率。设类 A 已有 m 个对象 $\{P_1, P_2, \dots, P_m\}$, 若存在一个对象 Q 要加入到类 A , 则必须满足:

$$\text{dist}(Q.a, v) \leq kS_D \quad (7)$$

式中, $Q.a$ 表示对象 Q 的非空间属性值; v 表示 $\{P_1.a, \dots, P_m.a, Q.a\}$ 的均值; S_D 表示类 $\{P_1.a, \dots, P_m.a, Q.a\}$ 的标准差,反映待划分对象与已知空间簇质心的偏离程度; $\text{dist}(Q.a, v)$ 表示空间对象与聚类集合之间的差异,采用欧氏距离进行度量,针对多维非空间属性要归一化处理。当对象 Q 满足式(7)时, Q 可以加入到类 A 中。

定义 11 聚类。给定一个聚类中心点 P , 根据属性变化率按照广度优先遍历方法搜寻该对象的直接邻居和间接邻居,直到没有新的对象加入,一个聚类簇形成。

定义 12 噪声。给定一个对象 P , 如果 P 不属于任何簇,认定 P 为噪声。

2 聚类阈值分析与算法实现

2.1 聚类阈值分析

如果 Sub_D_T 中非空间属性分布不均匀,采用全局参数聚类有可能将局部相似的对象划入噪声簇中或者其他簇中。以图 1(a) 数据为例,原始数据可分为两个簇 C_1 和 C_2 , 两簇的属性空间分布不均匀。若采用多约束聚类方法 [13] 可能得到两种结果: ① 只能发现簇 C_1 , 簇 C_2 被划分为噪声点, 如图 1(b) 所示; ② 当数据集中存在明显的异常值时, 簇 C_1 和簇 C_2 会合并成一个类, 如图 1(c) 所示。

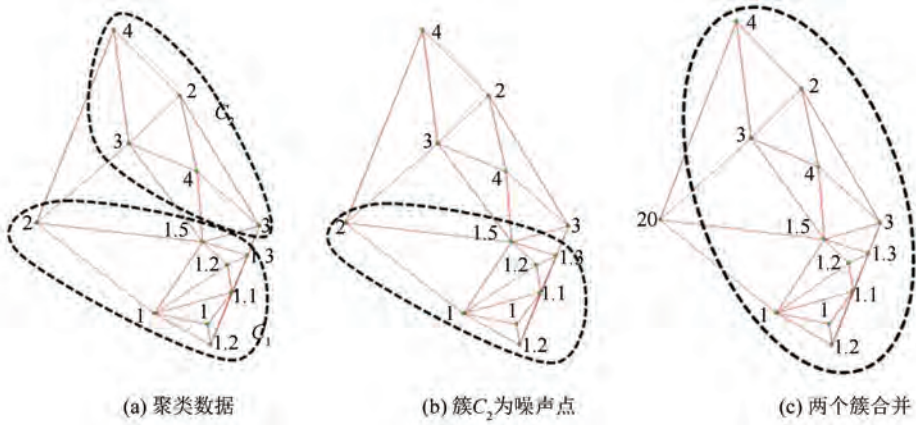


图1 全局参数聚类结果

Fig.1 Clustering Results Based on Global Parameter

本文提出的属性变化率能够描述对象间的相对变化信息,可以同时满足属性空间分布不均匀和均匀两种情况。此处 k 是控制参数,用于调节空间对象与聚类集合的偏离程度, k 值越大属性变化率的敏感性越低。为了保证参数 k 能够自动地适应不同的聚类需求,引入了最优分割指数(partitioning best method,PBM)^[17]。PBM 指数能够满足紧密性与分离性的聚类准则,指数越大聚类结果越可靠,具体为:

$$PBM = \left(\frac{1}{N_C} \frac{E_1}{E_{N_C}} D_{N_C} \right)^2 \quad (8)$$

$$E_{N_C} = \sum_{i=1}^{N_C} E_i, E_i = \sum_{j=1}^{N_i} \|x_j - v_i\| \quad (9)$$

$$D_{N_C} = \max \sum_{i,j=1}^{N_C} (\|v_i - v_j\|) \quad (10)$$

式中, N_C 表示簇的数量; N_i 表示簇 C_i 中实体数量; v_i 表示簇 i 的质心; E_i 表示簇 C_i 的内部距离(簇内所有空间对象到其质心的距离之和); E_1 表示数据集只分为一类的聚类内部距离; E_{N_C} 表示数据集分为 N_C 个簇的聚类内部距离; D_{N_C} 表示

空间簇间的分离度,随 N_C 增大而增大,最大值为数据集中最远两个簇的质心距离。

以图 2(a)的模拟数据为例阐述上述非空间属性聚类过程,图 2 中结点的属性密度值按照升序排序为 $(v_4, v_3, v_1, v_2, v_5, v_6, v_{12}, v_{11}, v_8, v_{10}, v_7, v_9)$,以 v_4 为聚类中心点按照 $\rho(P)$ 升序访问 v_4 的直接邻居 $\{v_3, v_1, v_5, v_2, v_6\}$,PBM 值最大时属性变化率中 k 取 1.8, $\{v_4, v_3, v_1, v_5, v_2, v_6\}$ 可以形成一个初始簇,将 $\{v_3, v_1, v_5, v_2, v_6\}$ 作为初始点按照上述方式搜索符合阈值条件的邻居,最终得到一个簇,如图 2(b)所示。聚类 1 为 $\{v_4, v_3, v_1, v_5, v_2, v_6, v_8\}$ 在剩余未聚类的对象中,以 v_{12} 为聚类中心点,按照广度优先遍历的方法聚类,得到最后的结果,如图 2(c)所示。聚类 2 为 $\{v_{12}, v_{13}, v_{10}, v_{11}, v_7, v_9\}$

2.2 算法实现

本文方法采用“先空间后属性”的聚类策略,具体步骤如下。

1) 构建Delaunay三角网并对其施加边长约

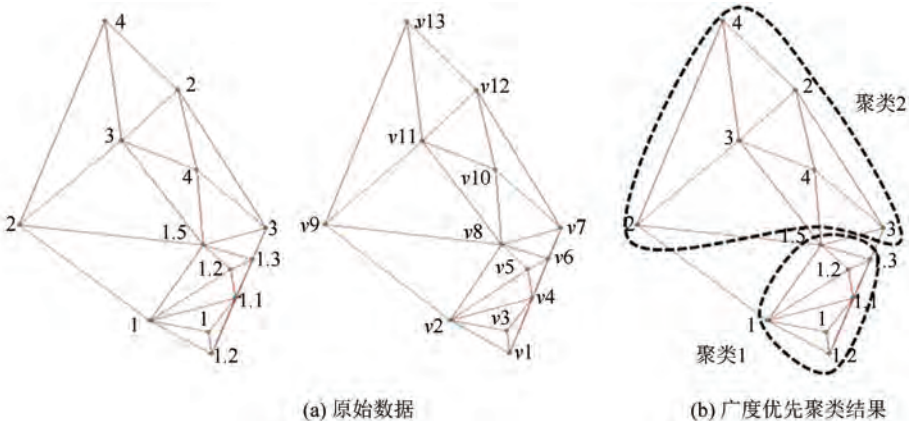


图2 基于属性密度的广度优先聚类结果

Fig.2 Clustering Result Based on Breadth First Traverse

束,删除不一致边,得到对象间的空间邻近关系。

2)计算每个对象的属性密度,对 Delaunay 三角网内每个对象属性密度进行升序排序。

3)非空间属性聚类。首先选取属性密度最小值作为初始聚类中心;然后按照广度优先遍历方法搜索该对象的直接邻居,将符合阈值的对象聚类合并,并标识为已聚类;依次循环搜寻该对象的间接邻居,向外不断扩散,将符合阈值的对象加入簇中,直到没有新对象加入,一个聚类簇形成。最后对于未被标识聚类的对象,迭代步骤 1)~3),遍历完所有对象,聚类结束,没有被标识到任何一类簇的对象视为噪声。

3 实验例证

本文以城市商业中心提取为例来验证方法的可行性,利用 Visual Studio 2010 C# 环境结合 ArcEngine 组件二次开发实现。采用的数据为 2011 年扬州市中心城区商业用地现状图,商业用地地图斑数为 735 个,如图 3(a)所示。城市地价直接决定不同商业类型的选址和布局^[18,19],本文以扬州市的商业用地价格为属性数据进行实证研究。

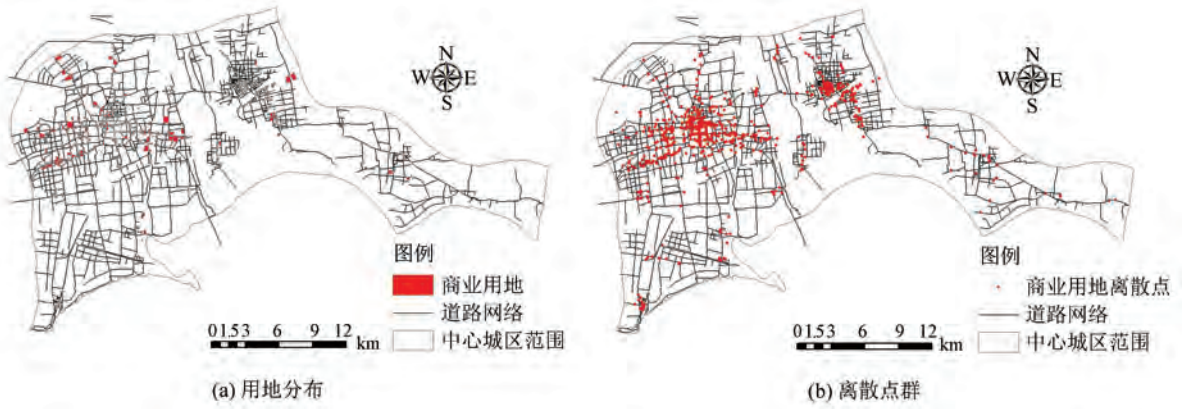


图 3 扬州市商业用地分布情况及商业用地离散点群

Fig.3 Commercial Land Distribution and its Discrete Points of Yangzhou City

首先将面状单元的商业地块抽象成离散点群(图 3(b))并构建 Delaunay 三角网,对三角网进行整体边长约束和局部边长约束($\gamma=0.4638$),

其具体过程如图 4(a)、4(b)所示,得到空间位置聚类结果(图 4(b))。由于篇幅限制,图 4(a)和图 4(b)给出了聚类结果的主体部分。

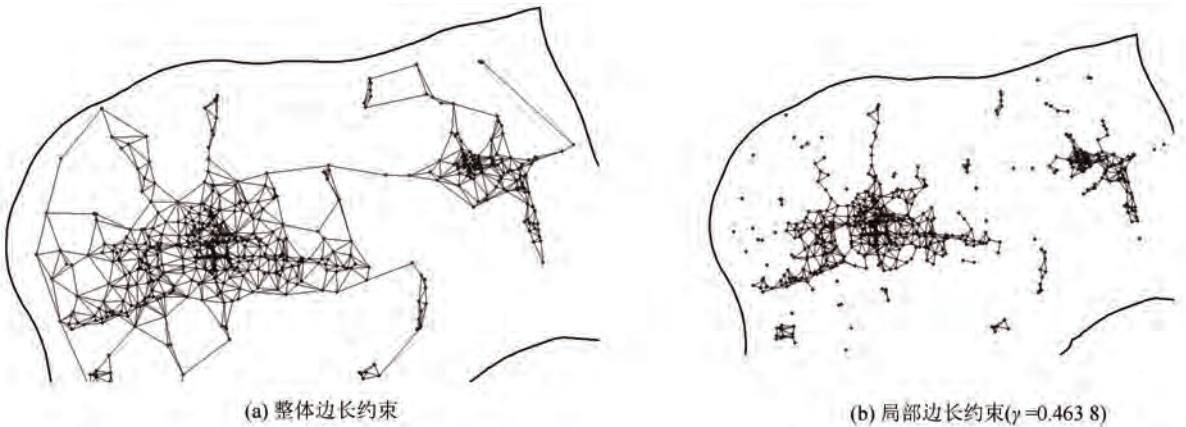


图 4 顾及空间位置的 Delaunay 三角网聚类过程

Fig.4 Delaunay Triangulation Clustering Based on Spatial Proximity in Commercial Discrete Points

将商业用地价格归一化到 $[0,10]$ 区间内,采用多约束聚类方法^[13]和本文方法($k=1.4, k=1.7, k=2.0$)分别进行非空间属性聚类,得到图 5 的聚类结果。

从图 5 可以看出,当属性变化率 $k=1.4$ 时,聚类数目、均值同多约束方法相似,两者聚类效果整体一致,也证明了本文方法的有效性。但多约束方法的属性阈值采用的是一种全局参数,导致

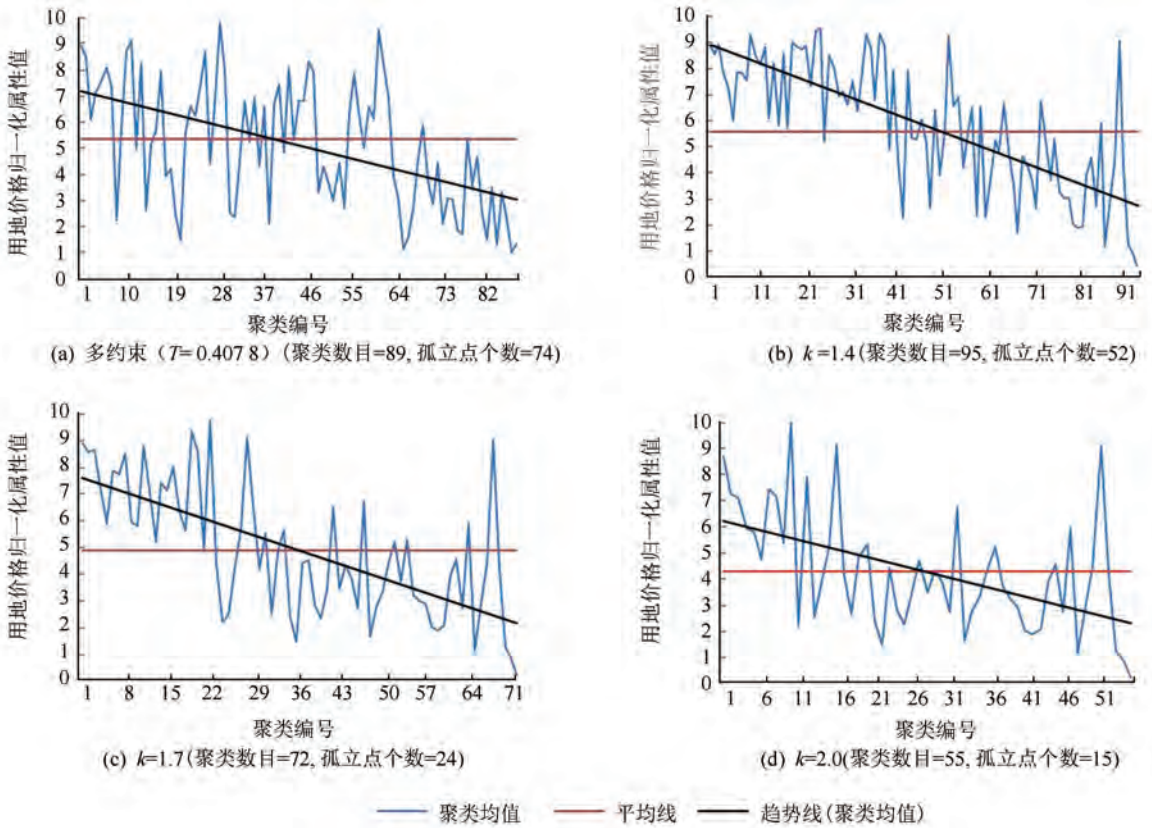


图5 不同方法的聚类结果对比

Fig.5 Comparison of the Clustering Results Using Different Methods

若干相似的簇无法识别,被划分成孤立点或噪声,因此其孤立点个数是最多的;当 $k=1.4$ 时,聚类阈值较小,聚类数目也会增多,可能导致簇间分离性不明显;随着 k 值不断变大,当 $k=2.0$ 时,聚类阈值较大,聚类数目也会减少,但这有可能会使簇内对象属性值出现较大差异。为了确定合适的 k 值,采用 PBM 值评价出最可靠的聚类结果,图 6 显示了不同 k 值 PBM 的取值情况。

7.063 4~9.797 3 界定为扬州市商业中心,如图 7 (a)所示,黑色圆圈即为聚类的城市商业中心。

1996 年扬州市城市中心体系规划包括两个城市主中心(文昌阁商圈与河东商务中心),一个城市副中心(西部副中心),将城市商业中心提取结果与扬州市城市空间格局(1996~2010)(图 7 (b))进行对比分析,区域 1、2、3 在 1996~2010 年间稳步发展,已形成一定的商业规模。另外,2011 年扬州市进行了行政区划调整,江都撤市并区,图 7(a)中提取的商业中心 4 表示江都城区中心。城市商业中心的识别可以为城市空间结构分析和未来城市规划提供必要的信息支持。

4 结 语

本文方法能够顾及实体间空间位置与非空间属性相似性的局部变化信息,更客观、准确地识别非空间属性相似的簇。属性密度、属性变化率的定义和广度优先聚类过程都兼顾了属性空间分布不均的情况,能够适应复杂的属性空间分布情况和输出相对稳定的聚类结果。与现有的 Delaunay 三角网空间聚类方法相比,本文方法具有在属性空间分布不均匀的情况下能够顾及属性分布

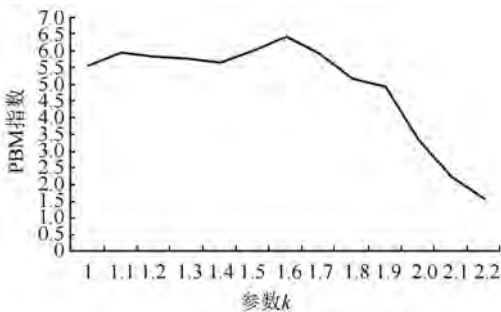
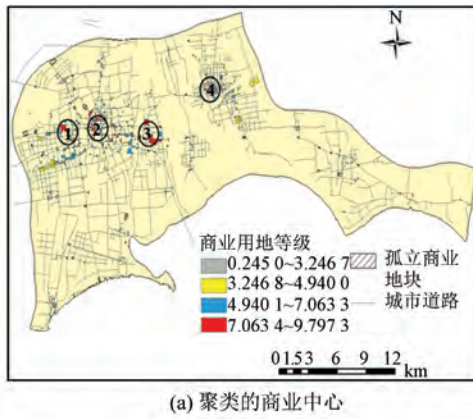


图6 不同 k 值 PBM 取值

Fig.6 PBM Value With the Change of Kvalue

当 k 取 1.6 时,PBM 值最大,计算出 $k=1.6$ 时每个聚类的均值,据此将扬州市商业用地分为 4 个等级:0.245 0~3.246 7,3.246 8~4.940 0,4.940 1~7.063 3,7.063 4~9.797 3,将分级范围



(a) 聚类的商业中心



(b) 城市空间格局

图 7 扬州城市商业中心提取结果对比

Fig.7 Comparison Between Result of City Commercial Centers Extraction and Actual Situation in Yangzhou

的局部信息,满足属性空间分布不均匀和均匀两种情况的优点。本文方法采用 PBM 值引导参数 k 的选取,可以自动地适应不同聚类需求,得到可靠的聚类结果。进一步的研究内容主要集中在高维数据相似性度量及其空间聚类方法上。

参 考 文 献

- [1] Deng Min, Liu Qiliang, Li Guangqiang, et al. Spatial Clustering Analysis and its Application[M]. Beijing: Science Press, 2011; 2-5 (邓敏,刘启亮,李光强,等.空间聚类分析及应用[M].北京:科学出版社,2011;2-5)
- [2] Kang I S, Kim T, Li K J. A Spatial Data Mining Method by Delaunay Triangulation[C]. The 5th International Workshop on Advances in Geographic Information Systems, Las Vegas, USA, 1997
- [3] Okable A, Boots B, Sugihara K, et al. Spatial Tessellations: Concepts and Applications of Voronoi Diagrams[M]. Hoboken: John Wiley & Sons, Inc, 2009
- [4] Eldershaw C, Hegland M. Cluster Analysis Using Triangulation[J]. *Computational Techniques and Applications*, 1997: 201-208
- [5] Kolingerová I, Alik B. Reconstructing Domain Boundaries Within a Given Set of Points, Using Delaunay Triangulation[J]. *Computers & Geosciences*, 2006, 32(9): 1 310-1 319
- [6] Estivill-Castro V, Lee I. Multi-level Clustering and its Visualization for Exploratory Spatial Analysis [J]. *GeoInformatica*, 2002, 6(2): 123-152
- [7] Deng M, Liu Q, Li G, et al. Field-Theory Based Spatial Clustering Method[J]. *Journal of Remote Sensing*, 2010, 14(4):694-700
- [8] Zhou Cuizhu, Zhu Jianjun, Shi Yan. A Multi-level Spatial Clustering Based on Distance Constraints[J]. *Science of Surveying and Mapping*, 2014, 39(10): 98-101 (周翠竹,朱建军,石岩.一种基于双重距离约束的多层次空间聚类方法[J].测绘科学,2014,39(10):98-101)
- [9] Mundur P, Rao Y, Yesha Y. Keyframe-Based Video Summarization Using Delaunay Clustering[J]. *International Journal on Digital Libraries*, 2006, 6(2): 219-232
- [10] Deng Min, Peng Dongliang, Liu Qiliang, et al. A Hierarchical Spatial Clustering Algorithm Based on Field Theory[J]. *Geomatics and Information Science of Wuhan University*, 2011, 36(7): 847-852 (邓敏,彭东亮,刘启亮,等.一种基于场论的层次空间聚类算法[J].武汉大学学报·信息科学版,2011,36(7):847-852)
- [11] Jiao Limin, Hong Xiaofeng, Liu Yaolin. Self-organizing Spatial Clustering Under Spatial and Attribute Constraints[J]. *Geomatics and Information Science of Wuhan University*, 2011, 36(7): 862-866 (焦利民,洪晓峰,刘耀林.空间和属性双重约束下的自组织空间聚类研究[J].武汉大学学报·信息科学版,2011,36(7):862-866)
- [12] Shi Yan, Liu Qiliang, Deng Min, et al. A Hybrid Spatial Clustering Based on Graph Theory and Spatial Density[J]. *Geomatics and Information Science of Wuhan University*, 2012, 37(11): 1 276-1 280 (石岩,刘启亮,邓敏,等.融合图论与密度思想的混合空间聚类方法[J].武汉大学学报·信息科学版,2012,37(11):1 276-1 280)
- [13] Liu Qiliang, Deng Min, Shi Yan, et al. A Novel Spatial Clustering Method Based on Multi-constraints[J]. *Acta Geodaetica et Cartographica Sinica*, 2011, 4: 509-516 (刘启亮,邓敏,石岩,等.一种基于多约束的空间聚类方法[J].测绘学报,2011,4:509-516)
- [14] Liu Q, Deng M, Shi Y, et al. A Density-Based Spatial Clustering Algorithm Considering Both Spatial-

- proximity and Attribute Similarity[J]. *Computers & Geosciences*, 2012, 46: 296-309
- [15] Estivill-Castro V, Lee I. Argument Free Clustering for Large Spatial Point-Data Sets Via Boundary Extraction from Delaunay Diagram [J]. *Computers, Environment and Urban Systems*, 2002, 26 (4): 315-334
- [16] Liu D, Sourina O. Free-Parameters Clustering of Spatial Data with Non-uniform Density[C]. *Cybernetics and Intelligent Systems*, 2004 IEEE Conference on, Singapore, 2004
- [17] Pakhira M K, Bandyopadhyay S, Maulik U. Validity Index for Crisp and Fuzzy Clusters[J]. *Pattern Recognition*, 2004, 37(3): 487-501
- [18] Zhang Wenzhong. *Economic Location Theory* [M]. Beijing: Science Press, 2000; 100-107 (张文忠. *经济区位论* [M]. 北京: 科学出版社, 2000; 100-107)
- [19] Battaglia F, Borruso G, Porceddu A. Real Estate Values, Urban Centrality, Economic Activities. A GIS Analysis on the City of Swindon (UK) [M]. Heidelberg: Springer Berlin, 2010

A Spatial Clustering Method Based on Uneven Distribution of Non-spatial Attributes——Identifying City Commercial Center

ZHU Jie^{1,2} SUN Yizhong^{1,2} CHEN Lvvyu^{1,2} ZHOU Wei^{1,2} MENG Yaowei^{1,2}

1 Key Laboratory of Virtual Geographic Environment of Ministry of Education, Nanjing Normal University, Nanjing 210023, China

2 Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China

Abstract: Spatial clustering is an important tool for spatial data mining and spatial analysis. It is important for the clustering results in many applications to meet the requirement that spatial objects in the same cluster are similar in both the spatial and the attribute domains. To solve the problems of existing methods based on Delaunay triangulation, this paper proposes a spatial clustering method considering uneven distribution of non-spatial attributes. The proposed algorithm involves two main steps: the first is to construct spatial proximity relationships, and the second is to cluster spatial objects with similar attributes. Delaunay triangulation with edge length constraints is first employed to construct spatial proximity relationships among objects. To obtain satisfied results in spatial clustering with attribute similarity, the breadth first traverse (BFT) clustering algorithm is used in accordance with variation rate of attribute to adapt to the local change information of attribute distribution. The performance of the proposed algorithm was evaluated experimentally through comparison with one of the leading state-of-the-art alternatives: multi-constraints algorithm. The results show that our method outperforms the comparative algorithm as attributes are unevenly distributed in space, and provides a quantitative research method in city commercial center extraction. The effectiveness and practicability of the proposed algorithm illustrates three advantages of our algorithm: i) it can reflect the tendency of the entity attribute in the spatial distribution; ii) it can meet the requirement that attributes are unevenly distributed in space; iii) it can discovery clusters with arbitrary shape and is robust to outliers.

Key words: spatial clustering; Delaunay triangulation; uneven distribution of non-spatial attributes; attribute threshold; city commercial center

First author: ZHU Jie, PhD candidate, specializes in spatial data representation and mining. E-mail: Chu_Je@163.com

Corresponding author: SUN Yizhong, PhD, professor. E-mail: sunyizhong_cz@163.com

Foundation support: The National Natural Science Foundation of China, Nos.41671392,41371374,41301405; Special Program for Basic Research of Sci-tech Police of Ministry of Public Security, No. 2015GABJC39.