文章编号:1671-8860(2012)11-1276-05　　**文献标志码**:A

融合图论与密度思想的混合空间聚类方法

石　岩[1]　刘启亮[1]　邓　敏[1]　林雪梅[1]

(1　中南大学地球科学与信息物理学院,长沙市麓山南路 932 号,410083)

摘　要:提出了一种融合图论与密度思想的空间聚类方法——HGDSC。该方法首先借助附加约束的 Delaunay 三角网来建立空间实体之间的邻接关系,然后对基于密度的聚类方法进行改进,顾及空间邻近与非空间属性相似性进行聚类。特别地,该方法只需要一个输入参数。模拟数据和实际数据验证表明,HGDSC 方法能够发现任意形状和密度变化的空间簇,并且可以很好地识别噪声点。

关键词:空间聚类;非空间属性;Delaunay 三角网;密度

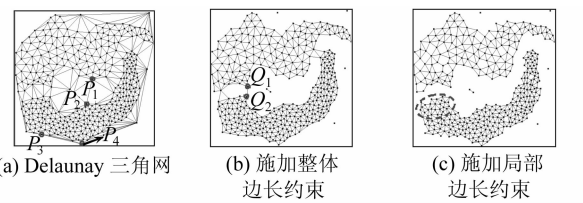
中图法分类号:P208

目前,空间聚类已在实际中得到了广泛的应用,如遥感图像分析[1,2]、地震分析[3]等。现有的空间聚类方法可以分为以下两类[4]:① 只顾及空间属性的空间聚类方法;② 同时顾及空间属性和非空间属性的空间聚类方法。为了更好地挖掘空间数据库内实体间的关联规则与分布模式,空间聚类方法需要能够同时顾及空间实体之间的邻近性与非空间属性之间的相似性[5]。近年来,一些学者对顾及非空间属性的空间聚类方法进行了初步的研究[2,5-10]。现有顾及非空间属性的空间聚类方法的主要局限性在于:① 基于全局的参数设置(如 Eps、MinPts)难以适应复杂的空间分布模式,如密度变化的情况;② 为了顾及非空间属性相似而引入过多的附加参数,从而影响了空间聚类方法的可用性。针对这些问题,本文充分借助了基于图论与基于密度两种空间聚类方法的优势,发展了一种新的混合聚类方法(a hybrid spatial clustering method based on graph theory and spatial density, HGDSC)。

1　混合空间聚类方法——HGDSC

1.1　空间邻近域的构建

空间邻近域的构建是 HGDSC 方法的首要步骤。现有的空间聚类方法构建空间邻近域的方法主要有 Eps 法[1]和 KNN[11]法。其中,Eps 法难以适应密度变化较大的空间数据,而 KNN 法在 K 值选择方面增大了用户的使用难度。Delaunay 三角网[12]是建立空间实体间邻接关系的一个有力手段,能够有效描述实体间的拓扑关系,但对于分布不规则的数据集,Delaunay 三角网建立的邻接关系在边缘处存在一定的误差,如图 1(a)中边 P_1P_2 和 P_3P_4,原因是 Delaunay 三角网构建的实体间的邻接关系缺乏距离约束。为此,本文采用整体与局部相结合的层次策略分别对三角网的边长进行修剪,并发展了整体与局部边长约束准则。



(a) Delaunay 三角网　(b) 施加整体边长约束　(c) 施加局部边长约束

图 1　构建空间邻近域

Fig. 1　Construction of the Spatial Neighborhood

定义 1　整体边长约束准则。对于 Delaunay 三角网中的任意实体 P_i,整体边长约束准则记为 Global_Distance_Constraint(P_i),表达为:

$$\text{Global\_Distance\_Constraint}(P_i) = \text{Global\_Mean}(\text{DT}) + \frac{\text{Global\_Mean}(\text{DT})}{\text{Local\_Mean}(P_i)} * \text{Global\_SD}(\text{DT}) \tag{1}$$

收稿日期:2012-09-15。

项目来源:国家 973 计划资助项目(2012CB719906);国家教育部新世纪优秀人才基金资助项目(NECT-10-0831);江苏省资源环境重点实验室开放基金资助项目(JS201101);中南大学前沿研究计划资助项目(2010QYZD002)。

式中,Global_Mean(DT)表示 Delaunay 三角网所有边长的平均值;Global_SD(DT)表示 Delaunay 三角网所有边长的标准差;Local_Mean(P_i)表示空间实体 P_i 的邻域里所有边长的平均值;$\frac{\text{Global_Mean(DT)}}{\text{Local_Mean}(P_i)}$表示调节系数。

在 Delaunay 三角网中,如果与 P_i 直接连接的边长度大于或等于 Global_Distance_Constraint(P_i),则首先进行删除,如图1(b)所示。删除整体长边后,仍然存在一些局部长边,如图1(b)中的边 Q_1Q_2,因此需要进一步施加局部边长约束。

定义 2　局部边长约束准则。图 G_i 为施加整体边长约束后得到的任一子图,P_j 为 G_i 中的一个顶点,则其局部边长约束准则记为 Local_Distance_ Constraint(P_j),表达为:

$$\text{Local\_Distance\_Constraint}(P_i) = \text{Local\_Mean}^2(P_i) + 2 * \sum_{j=1}^{n} \text{Local\_SD}(P_j)/n, P_j \in G_i \quad (2)$$

式中,Local_Mean^2(P_i)表示 P_i 的 2 阶邻域内所有边长的平均值;Local_SD(P_j)表示图 G_i 中与 P_j 直接连接边的标准差。

在任一子图 G_i 中,针对每个顶点 P_i 2 阶邻域内的所有边,若其长度大于或等于 Local_Distance_Constraint(P_j),则进行删除,如图 1(c)所示,边界和空洞处的长边被有效删除。因此,可以根据施加整体边长约束和局部边长约束后的 Delaunay 三角网来定义空间邻近域。

定义 3　空间邻域。在任一施加整体和局部边长约束后获得的子图 C-DT$_i$ 中,对于任一空间实体 P_j,所有与 P_j 直接通过边连接的实体构成了 P_j 的空间邻域,记为 NN(P_j)。

不难发现,依据上述方法构建的空间邻近域能够适应空间数据的密度差异,且避免了人为输入参数的干扰。

1.2　基于密度的非空间属性聚类

构建空间邻近域后,需要进一步顾及非空间属性进行聚类。Dist(P_1, P_2)表示空间实体 P_1 和 P_2 之间的非空间属性距离,本文采用欧氏距离(其中多维非空间属性首先需要进行归一化处理);T 表示非空间属性距离阈值。本文进一步扩展了基于密度的空间聚类思想,引入了空间间接可达与密度指数等新概念。下面给出几个重要的定义。

定义 4　空间直接可达。对于 C-DT$_i$ 中的任一实体 P_i,如果 $Q_i \in N(P_i)$,并且 Dist(P_i,Q_i)$\leqslant T$,则称 Q_i 和 P_i 空间直接可达。

定义 5　空间间接可达。对于一个空间实体集 CLU(其中 CLU 中的实体个数大于等于 2),如果实体 Q_i 满足以下两个条件,则称实体 Q_i 与集合 CLU 空间可达:① $Q_i \in N(P_i)$,且 $P_i \in$ CLU;② Dist(Q_i,Avg(CLU))$\leqslant T$,其中 Avg(CLU)为 CLU 中所有实体的非空间属性值的平均值。

定义 6　密度指数。对于图 C-DT 中的实体 P_i,用 DI(P_i)表示 P_i 的密度指数,定义如下:

$$\text{DI}(P_i) = N_{\text{sdr}}(P_i) + N_{\text{sdr}}/|\text{NN}(P_i)| \quad (3)$$

式中,N_{sdr}(P_i)表示与实体 P_i 空间直接可达的实体个数;|NN(P_i)|表示实体 P_i 的空间邻域实体个数。

定义 7　空间聚类核。在未进行聚类的所有实体中,密度指数最大的空间实体称为空间聚类核。若最大密度指数的实体不止一个,则与其相应的邻域实体之间平均非空间属性差异最小的实体首先选为空间聚类核。

定义 8　扩展核。对于图 C-DT$_i$ 中的实体 P_i,若在其邻域 NN(P_i)中至少有一个实体与 P_i 空间直接可达,则称 P_i 为一个扩展核。

定义 9　空间相连。对于空间聚类核 O,P 与 Q 均与 O 所在的空间实体集合满足空间间接可达,且与 O 所在的空间实体集合中至少有一个空间扩展核满足空间直接可达,则称 P 与 Q 空间相连(P 与 Q 空间不邻近)。

定义 10　边界点。针对任一空间实体 P,其不能作为空间聚类核或扩展核的空间实体,但与某一邻近空间实体 Q 满足空间直接可达,并与 Q 所在的空间实体集合满足空间间接可达,则称 P 为一个边界点。

定义 11　空间簇。选定一个空间聚类核 O,所有与其满足空间直接可达、空间间接可达的扩展核、边界点构成一个空间簇。

1.3　算法描述

对于一个包含 N 个实体的空间数据库 SDB,给定一个非空间属性阈值 T,根据给出的定义,HGDSC 算法主要包含以下 3 个主要步骤。

1) 建立空间邻接关系,具体包括以下三个操作:① 建立 Delaunay 三角网,时间复杂度约为 $O(N\lg(N))$;② 施加整体边长约束准则,时间复杂度与 N 成线性关系;③ 施加局部边长约束准则,时间复杂度与 N 成线性关系。

2) 针对每个实体计算其密度指数,并进行递减排序,时间复杂度约为 $O(N\lg(N))$。

3) 顾及非空间属性进行空间聚类,具体包括以下几个操作:① 选择一个空间聚类核 P_i,在其

邻域 NN(P_i)中,对未聚类的扩展核根据密度指数进行排序;② 在 NN(P_i)中,将同时满足空间直接可达和空间间接可达的扩展核按照密度指数从大到小的顺序与 P_i 聚到一起,形成初始簇;③ 以加入到初始簇里的扩展核为新的中心,按照①和②的策略继续扩展;④ 当没有实体可以加入到以 P_i 为空间聚类核的簇时,一个空间簇形成;⑤ 迭代进行①~④操作,当所有实体都遍历时,聚类结束,没有加入到任何簇的实体标识为噪声点。

步骤 3)的时间复杂度主要源自两方面:① 对每个实体的邻域根据密度指数进行排序。Delaunay 三角网中每个实体的邻域点个数的平均值约为 6,所以排序的时间复杂度小于 $O(6\lg 6N)$;② 聚类的扩展与 N 成线性关系,因此,HGDSC 算法总的时间复杂度约为 $O(N\lg(N))$。

由上述聚类步骤可以发现,HGDSC 方法在基于密度聚类方法的基础上进行了两方面的重要扩展:① HGDSC 方法同时顾及了空间实体整体与局部的差异。空间直接可达的定义度量了两个直接邻近实体的相似性,而空间间接可达的定义度量了单个实体与聚类集合间的相似性,这种策略可以有效顾及空间数据分布的趋势性与渐变特性;② HGDSC 方法聚类是有序的,密度指数的定义同时兼顾了"纯度"的概念。传统基于密度的聚类方法是不考虑聚类次序的,数据输入次序的差异常导致不同的聚类结果。HGDSC 方法中,依据了"密度"最大的部分最先进行聚类的原则对聚类次序进行约束。空间聚类核的定义用来选取"密度"最大的区域开始聚类,每次合并实体时,也是按照密度最大的原则进行,因此,HGDSC 方法可以有效避免聚类次序导致最终结果的多样性。

2 实验分析及应用

本文设计了两个实验。实验一采用两组模拟数据 SDB1 和 SDB2(如图 2 所示)模拟了两组一维非空间属性,实验结果分别与一体化方法、GDBSCAN 及 GeoSOM 进行了比较;实验二采用 2009 年我国陆地区域 601 个气象站点的年平均气温数据。非空间属性阈值 T 均设为最邻近实体非空间属性差异的平均值[5]。GDBSCAN 算法中,Minpts 和 k 均设为 4[2]。

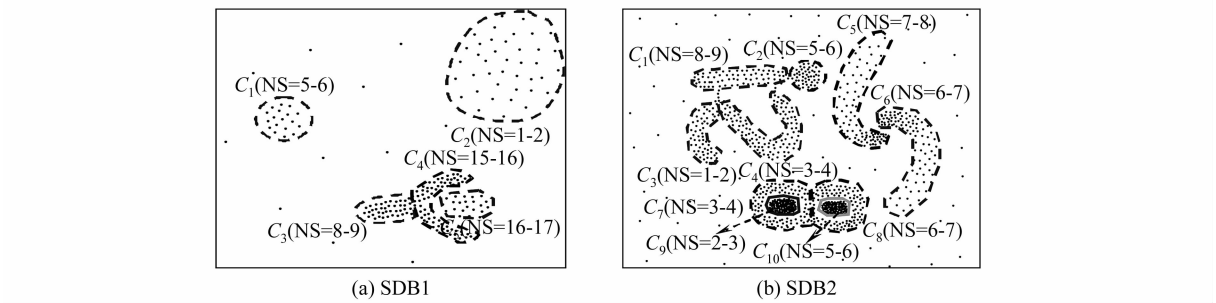


图 2 两组模拟数据集

Fig. 2 Two Simulated Datasets

2.1 模拟实验

两个模拟数据集 SDB1 和 SDB2 分别在文献[1,4]中采用的模拟数据的基础上添加了一维非空间属性(非空间属性值在一定约束范围内随机生成)。每个簇的边界以及簇内实体的非空间属性值(简称为 NS)范围已经进行了标注。SDB1 中有 5 个不同密度的簇,虽然 C_3 和 C_4 密度相近,但其非空间属性差异明显大于非空间属性的差异阈值 T。SDB2 包含 10 个簇,其中包含了密度不同的空间簇(如 C_5 和 C_9)、内部密度不均匀的簇(如 C_5 和 C_6)、空间邻近的空间簇(如 C_7 和 C_8)。此外,两个模拟数据中都包含了任意形状的空间簇及噪声点。

图 3 给出了 4 种方法对 SDB1 的聚类结果,可以发现:① HGDSC 方法准确识别了所有的空间簇和噪声点;② 一体化方法与 GeoSOM 方法难以发现任意形状的空间簇,且难以保证空间邻近的要求;③ GDBSCAN 方法难以适应密度的差异。

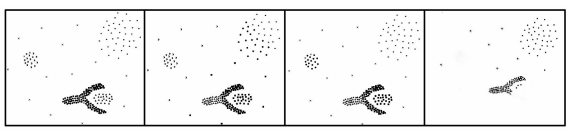
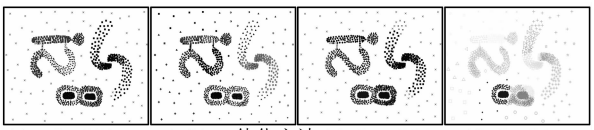


图 3 SDB1 聚类结果

Fig. 3 Clustering Results of SDB1

图 4 给出了 4 种方法对 SDB2 的聚类结果,可以发现,仅有 HGDSC 方法能够对预设的空间簇进行准确识别;其他三种方法均不能得到较好的结果,只有 GDBSCAN 能准确识别出 4 个簇。一体化方法一方面继承了划分方法的不足,难以

发现复杂形状的空间簇,另一方面定权加较为困难,故聚类效果不够理想。GeoSOM 方法实际上也类似于传统的划分方法,故难以发现任意形状的空间簇。GDBSCAN 算法采用全局的空间邻域构建策略,难以适应空间分布的分异特性。



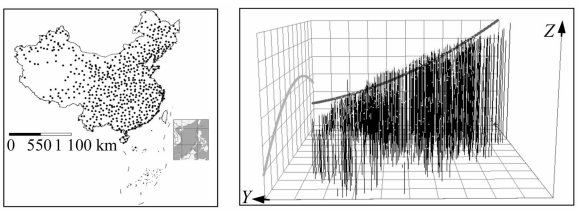
(a) HGDSC(T=0.87) (b) 一体化方法 (c) GDBSCAN (d) GeoSOM

图 4　SDB2 聚类结果

Fig. 4　Clustering Results of SDB2

2.2　实际应用

本文采用的实际数据集由中国气象局提供,包括了 1960～2009 年中国陆地区域 601 个气象站点的气温数据,气象站点分布如图 5(a)所示。对 2009 年的年平均气温数据进行趋势分析,如图 5(b)所示,发现从北到南有气温逐渐上升的趋势。由于对该数据具有一定的先验知识,因此,可以对聚类结果的有效性进行评价。

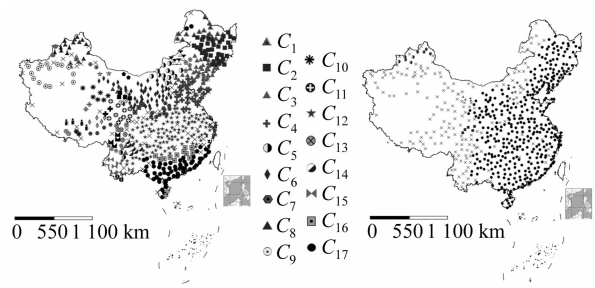


(a) 中国气象站点的分布　　(b) 2009 年年平均气温分布趋势

图 5　实际数据

Fig. 5　Real Data

HGDSC 与 GDBSCAN 方法的聚类结果如图 6 所示。从聚类结果可以发现:① HGDSC 方法将我国划分为 17 个气温特征区域,即 17 个主要的空间簇,每个簇的平均气温和标准差如表 1 所示,并且簇之间的差异较为明显,而簇内的标准差较小;② 自北向南分布有 10 个主要的簇 C_1～C_{10},这些簇之间不仅温度差异大,并且有递增的趋势,这与实际情况一致。同时可以发现,聚类结果能有效识别我国东部陆地区域的热带(C_{17})、亚热带(C_{12}与 C_{16})、温带(C_2、C_3、C_4)与寒带(C_1),同时从定量的角度对其进行了区分,可以为我国气象与气候学研究提供有益的参考。而 GDBSCAN 的聚类结果为自北向南的一个大簇,这与实际情况是相违背的,原因在于虽然整体上自北向南温度差异明显,但相邻站点的差异不大,整个温度增长态势是渐变的,而非突变。本文的聚类结果对于进一步研究我国气温的时空演变规律及突变特性具有重要的参考价值。



(a) HGDSC(T=2.2℃)　　(b) GDBSCAN (Eps1=180 km,Eps2=2.2℃)

图 6　HGDSC 和 GDBSCAN 聚类结果

Fig. 6　Spatial Clustering Result of Temperature Data by HGDSC and GDBSCAN

表 1　HGDSC 方法聚类结果统计信息/(℃)

Tab. 1　Statistical Information of the Clustering Result/(℃)

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}	C_{17}
实体数	16	40	38	30	13	53	11	9	15	7	11	133	6	7	7	12	57
平均值	0.4	4.7	8.7	13.4	5.1	9.3	13.4	4.8	12.7	4.9	2.2	16.8	8.7	13.1	19.1	19.6	22.2
标准差	0.9	1.2	1.3	0.7	1.0	1.0	0.7	0.8	0.9	1.2	1.2	1.2	1.2	1.1	1.1	0.4	1.1

3　结　语

本文提出了一种基于图论与密度的混合空间聚类方法——HGDSC,模拟实验和实际应用表明,HGDSC 方法能有效地识别任意形状、密度不同的空间簇以及噪声点,且需要较少的输入参数,效率高;时间复杂度约为 $O(N(\lg N))$,N 为空间点个数。进一步的工作将主要集中在空间聚类结果的定量评价以及研究高维非空间属性的空间聚类方法上。

参　考　文　献

[1] Ester M, Kriegel H P, Sander J, et al. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise[C]. The 2nd International Conference on Knowledge Discovery and Data Mining, Portland, USA, 1996

[2] Sander J, Ester M, Kriegel H P, et al. Density-based Clustering in Spatial Database: the Algorithm GDBSCAN and Its Applications[J]. Data Mining and Knowledge Discovery, 1998, 2(2): 169-194

[3] Pei T, Zhou C H, Zhu A X, et al. A New Ap-

proach to the Nearest-neighbour Algorithm to Discover Cluster Features in Overlaid Spatial Point processes[J]. International Journal of Geographic Information Science, 2006, 20(2): 153-168

[4] Deng M, Liu Q L, Cheng T, et al. An Adaptive Spatial Clustering Algorithm Based on Delaunauy Triangulation[J]. Computers, Environment and Urban Systems, 2011, 35(4): 320-332

[5] 李光强，邓敏，程涛，等. 一种基于双重距离的空间聚类算法[J]. 测绘学报，2008，37(4)：171-186

[6] 邓敏，刘启亮，李光强，等. 一种基于似最小生成树的空间聚类算法[J]. 武汉大学学报·信息科学版，2010，35(11)：1 360-1 364

[7] 刘启亮，李光强，邓敏. 一种基于局部分布的空间聚类算法[J]. 武汉大学学报·信息科学版，2010，35(3)：373-377

[8] Fernando B, Victor L, Marco P. The Self-organizing Map, the Geo-SOM, and Relevant Variants for Geosciences[J]. Computers & Geosciences, 2005, 31(2): 155-163

[9] Lin C, Liu K, Chen M. Dual Clustering: Integrating Data Clustering Over Optimization and Constraint Domains[J]. IEEE Transaction on Knowledge and Data Engineering, 2005, 17(5): 628-637

[10] 李新运，郑新奇，闫弘文. 坐标与属性一体化的空间聚类方法研究[J]. 地理与地理信息科学，2004，20(2)：38-40

[11] 李光强，邓敏，刘启亮，等. 一种适应局部密度变化的空间聚类方法[J]. 测绘学报，2009，38(3)：255-263

[12] Tsai V J D. Delaunay Triangulations in TIN Creation: an Overview and a Linear-time Algorithm[J]. International Journal of Geographical Information Systems, 1993, 7(6): 501-52

第一作者简介：石岩，硕士，主要从事空间聚类分析及其应用研究。

E-mail：shiyan0401060322@126.com

A Hybrid Spatial Clustering Method Based on Graph Theory and Spatial Density

SHI Yan[1] *LIU Qiliang*[1] *DENG Min*[1] *LIN Xuemei*[1]

(1 School of Geosciences and Info-physics, Central South University, 932 South Lushan Road, Changsha 410083, China)

Abstract: A hybrid spatial clustering method based on graph theory and spatial density (HGDSC) is developed. The HGDSC method employs Delaunay triangulation to model the spatial proximity relationships among spatial entities and the modified density-based clustering method, considering the similarity of both geometric distance and non-spatial attribute. Normally, the method can adapt to a spatial database which contains clusters of arbitrary shapes, non-homogeneous densities and/or large amount of noise. Only one input parameter is required. Experiments on both synthetic and real-world spatial dataset are utilized to demonstrate the effectiveness and advantages of the HGDSC method.

Key words: spatial clustering; non-spatial attribute; Delaunay triangulation; density

About the first author: SHI Yan, master, majors in spatial clustering analysis.

E-mail: shiyan0401060322@126.com