文章编号:1671-8860(2012)01-0096-05　　**文献标志码**:A

一种顾及障碍约束的空间聚类方法

石　岩[1]　刘启亮[1]　邓　敏[1]　王佳璆[1]

(1　中南大学测绘与国土信息工程系,长沙市麓山南路,410083)

摘　要:为了使得空间聚类分析更加适应实际情况,发展了一种同时顾及空间障碍约束与空间位置邻近的空间聚类方法。该方法采用 Delaunay 三角网描述实体间的邻近关系,并且不依赖用户指定参数。实验验证了本方法的有效性与优越性。

关键词:空间聚类;空间障碍;Delaunay 三角网;空间数据挖掘

中图法分类号:P208

空间聚类分析旨在将空间数据库实体划分为一系列具有一定意义的空间簇,簇内空间实体尽可能相似,簇间实体的差异性尽可能大,已经成功应用于地震空间分布模式分析[1]、公共设施选址[2]以及遥感图像分类[3]等众多领域。现有的空间聚类方法可大致分为以下几类:① 划分的方法[4-5];② 层次的方法[6-7];③ 基于密度的方法[8-10];④ 基于图论的方法[11-13];⑤ 基于格网的方法[14-15];⑥ 基于模型的方法[16-17]。然而,实际中空间实体间可能会存在一些空间障碍,如山脉、河流、桥梁等,导致实体间的通达性无法直接依据欧氏距离进行度量。

2000 年,Tung 等人[2]首次提出了顾及空间障碍的空间聚类问题——COE(clustering with obstacles entities)。现有顾及空间障碍的空间聚类[18-19]方法大致可分为以下 3 类:① 基于划分的方法;② 基于图论的方法[12,20];③ 基于密度的方法[21-22]。基于密度扩展的方法易受空间实体分布密度差异的影响,同时,聚类参数的选择涉及过多的先验知识。本文发展了一种顾及障碍的空间聚类新方法。

1　顾及障碍的空间聚类

借助 Delaunay 三角网空间聚类时,不一致边可以分为以下 3 类:① 过长的边,如簇与孤立点间的边;② 过短的边,如“链”上的边;③ 连接边,如两个簇形成的“颈”上的边。在空间数据分析中,一个空间过程是由整体上(大尺度)的确定变异与局部上(小尺度)的随机变异共同作用组成的[23]。因此,本文方法在聚类时首先打断整体上的长边,获得粗略的空间聚集划分;进而,考虑空间障碍的阻隔作用,打断与空间障碍相交的边;然后,顾及局部影响,打断局部的长边;最后,借助场论聚类[24]的思想,消除“颈”与“链”问题的影响,获得最终的聚类结果。“颈”是连接两个邻近簇的小部分实体,如图 1(a)下部两个球形簇之间的部分。“链”是由一系列噪声形成的呈线性结构的空间实体,如图 1(a)上部两个簇之间的空间实体。“颈”与“链”问题是空间聚类研究中的两个难点问题。

1.1　删除整体长边

建立 Delaunay 三角网后,首先从整体角度删除三角网的长边,形成粗略的聚集划分。Delaunay 三角网的平均边长和边长标准差能较好地体现空间实体分布的整体特征,故可以用于构造统计准则进行整体长边的删除。本文采用文献[11]中的类似策略对 Delaunay 三角网的整体长边进行删除。对于任意空间实体 p,e_i 表示与 p 直接邻接的边,则 p 对应的整体长边集合 Global_Long_Edges(p)可以定义为:

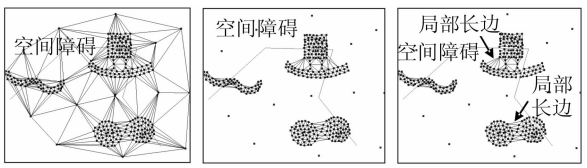
收稿日期:2011-09-29。

项目来源:国家 863 计划资助项目(2009AA12Z206);地理空间信息工程国家测绘局重点实验室开放研究基金资助项目(200916;201015);中南大学前沿研究计划资助项目(2010QYZD002)。

$$\text{Global\_LongEdges}(p)=\left\{e_i \,\middle|\, |e_i|>\text{Global\_Mean}+\text{Global\_SD}*\frac{\text{Global\_Mean}}{\text{Local\_Mean}(p)}\right\} \quad (1)$$

式中,Global_Mean 表示 Delaunay 三角网所有边长的平均值;Global_SD 表示 Delaunay 三角网所有边长的标准差;Locall_Mean(p)表示与 p 直接邻接的边长平均值。

针对每个空间实体,删除其对应的整体长边集合中的边。图 1(a)表示一个顾及空间障碍的聚类;图 1(b)表示删除整体长边后的子图,可见,孤立点与整体上比较明显的聚集部分能够很好地区分。



(a) 三角网与空间障碍　(b) 删除整体长边后的子图　(c) 顾及空间障碍后的子图

图 1　模拟数据集及删除整体长边、顾及空间障碍后的子图

Fig. 1　Subgraphs by the Removal of Global Long Edges and the Consideration of Spatial Obstacles

1.2　空间障碍叠置分析

删除整体长边后,需要进一步考虑空间障碍的影响。基于 Delaunay 三角网的聚类方法中,同一空间簇的实体通过三角网的边互相连接,当空间障碍存在时某些实体间的邻接关系将被隔断。因此,本文采取与 AUTOCLUST$^+$ 相同的策略,顾及空间障碍的影响,即如果空间障碍与 Delaunay 三角网的边相交,则将其打断。在实际操作中,两个图层(空间障碍层与空间点实体层)进行一次叠置分析操作,打断所有与障碍相交的边,可以进一步获得一系列的子图。如图 1(c)表示了顾及空间障碍后获得的子图,所有与障碍相交的边均进行了删除。这一策略可以简便、有效地顾及空间障碍的影响。进而,将考虑局部因素的影响。

1.3　删除局部长边

删除整体长边与顾及空间障碍后,某些局部长边依然存在,如图 1(c)所示。因此,需要发展相应的约束准则来删除这些局部的长边。针对一个空间实体,在其 2 阶邻域内考虑局部因素的影响,即仅考虑任一子图 G 中的一个顶点 p,到 p 的路径小于或等于 2 的所有顶点。对于任一子图 G_i(包含 n 个空间实体),其中任一空间实体 p,用 e_j 表示 p 的邻接边,则 p 对应的局部长边 Local_Long_Edges(p)可以定义为:

$$\text{Local\_Long\_Edges}(p)=\left\{e_i \,\middle|\, |e_i|>\text{Local\_Mean}^2(p)+\frac{\sum_{j=1}^{n}\text{Local\_SD}(p_j)}{n}, p_j\in G_i\right\} \quad (2)$$

式中,Local_Mean^2(p)表示 p 的 2 阶邻域内所有边长的平均值;Local_SD(p_j)表示图 G_i 中与空间实体 p_j 直接邻接边的标准差。

在子图中,针对每个空间实体删除其对应的局部长边,可得到进一步精化的子图。如图 2(a)所示,簇局部的长边被有效删除。式(2)实际上是传统极值统计准则的变种,但在 Delaunay 三角网中,每个实体的直接邻近实体数量较少(一般不超过 6 个),难以进行统计分析。采用实体的二阶邻域进行分析,一方面顾及了实体分布的局部特征;另一方面样本数量增加,降低了统计分析的误差。同时,采用直接邻接边方差的均值能够更好地反映实体空间分布的局部变化。

1.4　空间实体局部凝聚趋势提取

经过前面 3 个步骤的处理,虽然空间簇之间较长的不一致边得到了删除,然而空间簇之间形成的"颈"与"链"依然存在,如图 2(a)所示。删除局部短边的策略虽然可以部分解决"链"问题,但是还不够稳健且对于"颈"问题无法解决。本文借鉴场论聚类的思想[24]来解决这种"颈"问题和"链"问题。凝聚力可表达为:

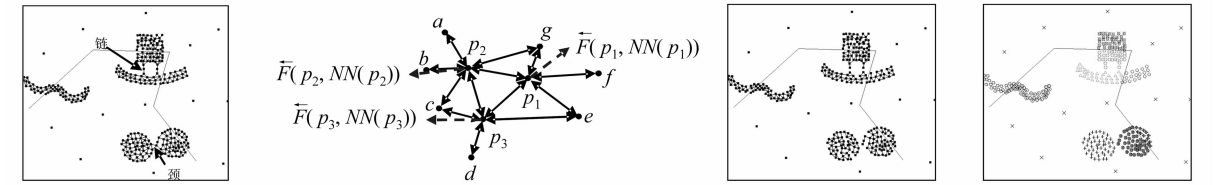
$$\vec{F}(p,q_i)=k\frac{1}{d^2(p,q_i)}m_p m_{q_i}\vec{e}_{p\to q_i}, q_i\in NN(p) \quad (3)$$

凝聚合力可表达为:

$$\vec{F}(p,NN(p))=\sum\vec{F}(p,q_i), q_i\in NN(p) \quad (4)$$

每个空间实体在凝聚合力的作用下,将有向对其凝聚力作用较大的实体聚集的趋势。依据文献[24]的定义,如果点 p 对 $NN(p)$ 中某个点 q_i 的引力方向与此合力的方向之间夹角小于 90°,那么则认为 p 有向 q_i 靠近的趋势,说明它们之间存在较强的邻接关系;反之,若夹角大于 90°则认为两者具有分裂的趋势,并打断实体间的边。基于这种思想,空间簇在其边缘处将形成自然的收缩趋势,邻接的空间簇其边缘处实体的凝聚合力方向具有较为明显的背离趋势,故可以有效解决"颈"和"链"问题。如图 2(b)所示,点 p_1 与点 e、f、g 有"靠近"趋势,与点 p_2 和 p_3 有"分裂"趋势;点

p_2与点a、b、c有“靠近”趋势，与点g、p_1和p_3有“分裂”趋势；点p_3与点p_2、c、d有“靠近”趋势，与点p_1和e有“分裂”趋势，故空间簇之间形成的“颈”可以有效区分，“链”问题的解决也与此类似。最后，每个通过Delaunay三角网的边连接的子图均视为一个空间簇，模拟的聚类结果如图3所示。



(a) 删除局部长边 (b) 局部凝聚力分析

图2 局部长边的删除及局部凝聚力分析

Fig. 2 Removal of Local Long Edges and Local Aggregation Force

(a) 空间实体局部凝聚趋势的提取 (b) 聚类可视化结果

图3 空间聚类结果

Fig. 3 Spatial Clustering Result

1.5 算法复杂度

1) 建立Delaunay三角网，复杂度约束为$O(n\lg(n))$；

2) 由于三角网中实体的邻接实体个数平均约为6，所以删除整体长边、删除局部长边、凝聚力分析的复杂度总和为$O(18n)$；

3) 分析Delaunay三角网的边是否与线障碍相交时，复杂度也是近似线性。

本文方法的整体复杂度约为$O(n\lg(n))$，运行效率较高，能够适应海量数据应用的要求。

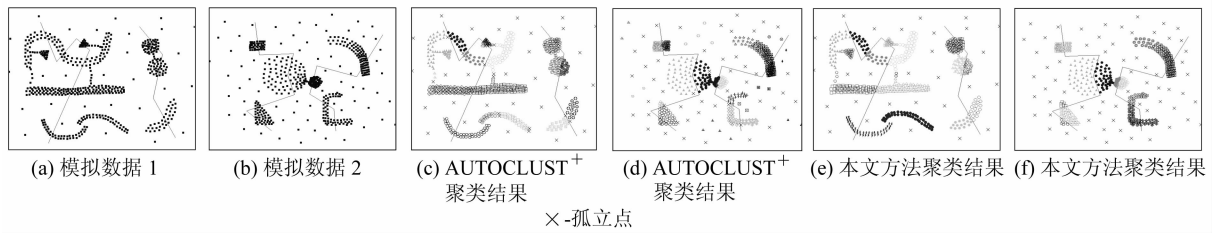
2 实验分析

本文设计了两组实验来证明本方法的有效性：实验一采用两组模拟数据，均在Arcgis9.2软件中模拟生成，实验结果与AUTOCLUST$^+$进行了比较；实验二采用华南某市的污染源数据，其中河流作为空间障碍，该数据源于我国华南某市环境保护规划项目(2006～2009)。

2.1 模拟实验

模拟数据如图4(a)和4(b)所示，AUTOCLUST$^+$的聚类结果如图4(c)和4(d)所示，本文方法的聚类结果如图4(e)和4(f)所示。

分析两种方法的聚类结果可以发现：① 本文方法可以有效地顾及空间障碍的影响，且能够识别复杂结构的空间簇，能有效避免孤立点、“颈”以及“链”的干扰；② AUTOCLUST$^+$方法虽然可以顾及空间障碍的影响，但是其聚类结果并不稳健，易受“颈”和某些短“链”的干扰，同时对于空间簇内部密度不均匀的情况过于敏感。进一步，将检验本文方法的实用性。



(a) 模拟数据1 (b) 模拟数据2 (c) AUTOCLUST$^+$聚类结果 (d) AUTOCLUST$^+$聚类结果 (e) 本文方法聚类结果 (f) 本文方法聚类结果

×-孤立点

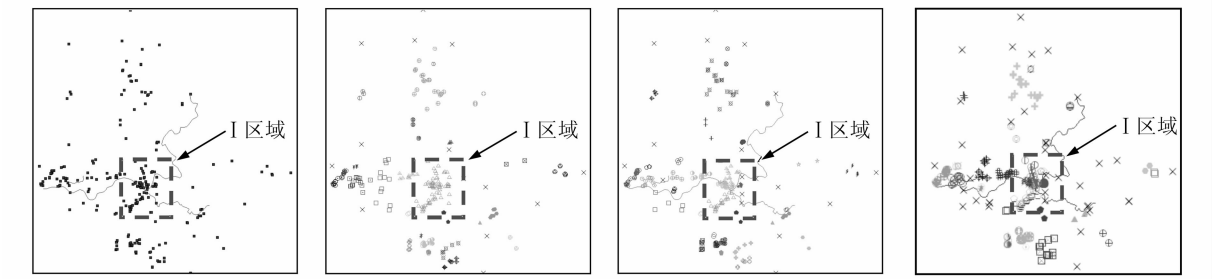
图4 模拟数据聚类结果

Fig. 4 Spatial Clustering Results of Simulated Datasets

2.2 实际应用

采用我国华南某市的污染源企业分布数据库来验证本文方法的实用性。图5(a)显示了污染源企业的空间分布，线实体表示该区域的主要河流，视为主要的空间障碍。数据来源于我国华南某市的环境保护规划项目(2006～2009)，其中一项内容即进行相关监测机构的空间位置选址，其一方面需要考虑污染源企业的集中分布特性，另一方面也要顾及空间可达性。本文采用顾及障碍约束的空间聚类方法来为这一问题提供辅助决策。图5(b)显示了本文方法不考虑空间障碍时的聚类结果；图5(c)显示了本文方法顾及河流障碍时的空间聚类结果；图5(d)给出了AUTOCLUST$^+$方法的聚类结果。分析本文方法的实验结果可以发现，当没有顾及河流障碍时，I区域作为一个大簇存在，虽然满足传统意义的空间集聚分布，但是由于河流障碍的存在，其内部的空间

可达性存在差异,可能给实际的监测工作带来一定的不便。而顾及障碍后,I 区域被划分为 4 个较小的空间簇,其内部的空间可达性较强。同时进一步分析这些空间簇的环境特征,对于环境保护中的污染控制以及关联分析将具有一定的指导意义。AUTOCLUST$^+$ 方法的聚类结果与本文方法基本一致,也从另一个方面验证了本文方法在实际应用中的可靠性。结合 §2.1 的实验内容,可以充分说明本文方法可适应更为复杂的空间聚类操作。



(a) 污染企业空间分布　(b) 空间聚类结果　(c) 本文方法聚类结果　(d) AUTOCLUST$^+$ 聚类结果

图 5　实际数据聚类结果(×-孤立点)

Fig. 5　Spatial Clustering Results of Real-life Datasets(×-outlier)

3　结　语

1) 本文方法可以有效地顾及空间障碍的影响,能够适应复杂的空间聚类操作;

2) 本文方法不需要人为输入参数,具有良好的自适应能力。

进一步地,将在实际应用中检验本文方法的适用性以及研究空间聚类结果的定量评价方法。

参　考　文　献

[1] Pei T, Zhu A X, Zhou C H, et al. A New Approach to the Nearest-neighbor Method to Discover Cluster Features in Overlaid Spatial Point Processes [J]. International Journal of Geographical Information Science, 2006, 20(2): 153-168

[2] Tung A K H, Hou J, Han J. COE: Clustering with Obstacles Entities, a Preliminary Study[C]. The 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Heidelberg, 2000

[3] 骆剑承, 梁怡, 周成虎. 基于尺度空间的分层聚类方法及其在遥感影像分类中的应用[J]. 测绘学报, 1999, 28(4): 319-324

[4] Macqueen J. Some Methods for Classification and Analysis of Multivariate Observations[C]. The 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, California, 1967

[5] Ng R, Han J. Efficient and Effective Clustering Method for Spatial Data Mining[C]. The 1994 International Conference on very Large Data Bases, Santiago, 1994

[6] Zhang T, Ramakrishnan R, Livny M. BIRCH: an Efficient Data Clustering Method for very Large Databases[C]. The ACM SIGMOD International Conference on Management of Data, Montreal, Canada, 1996

[7] Guha S, Rastogi R, Shim K. CURE: an Efficient Clustering Algorithm for Large Databases[C]. The 1998 ACM-SIGMOD International Conference on Management of Data, Seattle, Washington D C, 1998

[8] 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 the International Conference on Knowledge Discovery and Data Mining, Portland, OR, 1996

[9] Ankerst M, Breunig M M, Kriegel H P, et al. OPTICS: Ordering Points to Identify the Clustering Structure[C]. The 1999 ACM SIGMOD International Conference on Management of Data, Philadelphia, USA, 1999

[10] 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

[11] Estivill-Castro V, Lee I. Multi-level Clustering and Its Visualization for Exploratory Spatial Analysis [J]. GeoInformatica, 2002, 6(2): 123-152

[12] Estivill-Castro V, Lee I. AUTOCLUST: Automatic Clustering via Boundary Extraction for Mining Massive Point-Data Sets[C]. The Fifth International Conference on Geo-computation, New South Wales, Australia, 2000

[13] 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

[14] Wang W, Yang J, Muntz R. STING: a Statistical

Information Grid Approach to Spatial Data Mining [C]. The 1997 International Conference on very Large Data Bases, Athens, Greece, 1997

[15] Sheikholeslami G, Chatterjee S, Zhang A. Wave Cluster: a Multi-resolution Clustering Approach for very Large Spatial Databases[C]. The 24th International Conference on very Large Databases, New York, USA, 1998

[16] Dempster A, Laird N, Rubin D. Maximum Likelihood from Incomplete Data via the EM Algorithm [J]. Journal of the Royal Statistical Society, Series B, 1977, 39(1): 1-38

[17] Xu X, Ester M, Kriegel H P, et al. A Distribution-based Clustering Algorithm for Mining in Large Spatial Database[C]. The 14th International Conference on Data Engineering, Washington D C, USA, 1998

[18] Tung A K H, Hou J, Han J. Spatial Clustering in the Presence of Obstacles[C]. International Conference on Data Engineering(ICDE'01), Heidelberg, Germany, 2001

[19] Tung A K H, Han J, Lakshmanan L V S, et al. Constraint-based Clustering in Large Databases[C]. 2001 International Conference on Data Theory, Heidelberg, Germany, 2001

[20] Estivill-Castro V, Lee I J. AUTOCLUST+: Automatic Clustering of Point-Data Sets in the Presence of Obstacles[C]. The International Workshop on Temporal, Spatial and Spatial-Temporal Data Mining, Lyon, France, 2000

[21] Zaïane O R, Lee C H. Clustering Spatial Data When Facing Physical Constraints[C]. The IEEE International Conference on Data Mining(ICDM'02), Maebashi City, Japan, 2002

[22] Wang X, Rostoker C, Hamilton H J. DBRS+: Density-based Spatial Clustering in the Presence of Obstacles and Facilitators[OL]. ftp://cs.uregina.ca/Research/Techreports/2004-08.pdf, 2004

[23] Haining R. Spatial Data Analysis Theory and Practice[M]. London: Cambridge University Press, 2003

[24] 邓敏，刘启亮，李光强，等. 基于场论的空间聚类算法[J]. 遥感学报，2010，14(4)：702-709

第一作者简介：石岩，硕士生。主要从事空间聚类分析及其应用研究。

E-mail：shiyan0401060322@126.com

A Novel Spatial Clustering Method with Spatial Obstacles

SHI Yan[1] *LIU Qiliang*[1] *DENG Min*[1] *WANG Jiaqiu*[1]

(1 Department of Surveying and Geo-informatics, Central South University, South Lushan Road, Changsha 410083, China)

Abstract: Spatial clustering has been a major research field in spatial data mining; it aims to discover some useful patterns or outliers in a spatial database. In practice, spatial obstacles, as river or mountains should be fully considered in the process of spatial clustering. On that account, a novel spatial clustering method considering spatial obstacles is proposed in this paper. Delaunay triangulation is employed to model spatial proximate relations among entities, and the method can automatically discover clusters with complex structures without user-specified parameters. Experiments on both simulated database and real-world database are utilized to demonstrate the effectiveness and advantage of our method.

Key words: spatial clustering; spatial obstacle; Delaunay triangulation; spatial data mining

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

E-mail: shiyan0401060322@126.com