文章编号:1671-8860(2009)03-0335-04　　文献标志码:A

顾及距离与形状相似性的面状地理实体聚类

杨春成[1]　何列松[1]　谢　鹏[1]　周校东[1]

(1　西安测绘研究所,西安市雁塔路中段 1 号,710054)

摘　要:与点状地理实体不同,面状地理实体不仅具有位置特征,还具有形状特征。对于面状地理实体而言,仅考虑距离因素设计聚类准则是不全面的。综合考虑距离和几何形状相似性来设计聚类准则,实现了相应的聚类算法。实验证明,该算法适合面状地理实体的聚类分析。

关键词:空间聚类;面状地理实体;相似性准则

中图法分类号:P208

空间聚类问题可以用数学语言描述为:将 n 个空间对象$\{x_1,x_2,\cdots,x_n\}$划分为 c 个子集(聚类)$V_1,V_2,\cdots,V_c$,如果用 U_{ik} 表示对象 x_k 是否属于子集 V_i(第 i 个子集),就得到 n 个对象的一种划分,完成这种划分的操作就是空间聚类分析[1,2]。单一的聚类准则不能解决所有可能的聚类问题,因此人们提出了多种相似性函数,如最大似然准则、最大熵准则、最小体积准则、信息论准则和基于最小类内加权平方误差和准则等。基于距离平方和最小来设计相似性(或相异性)准则是常用的方法,许多聚类算法[3-5]都基于此实现。求 n 个对象中 c 个子集的距离平方和最小解是组合优化问题,文献[3,4]采用随机搜索算法来解决该问题,由于寻优次数受 Numlocal 参数的限制,这种随机搜索算法不能保证得到全局最优解,又因为寻找局部最优解的次数受 maxneighbor 参数的限制,算法也不一定能得到真正的局部最优解。另外,如果算法初值选择不当,算法最终得到的结果会受影响。文献[5]采用遗传算法来解决该问题,遗传搜索的结果是全局最优的,但在设计聚类准则时仅考虑了距离因素。

传统的聚类方法主要是针对点状地理空间实体设计的,而现实世界中的地理空间实体各式各样,按照几何定位特征和空间维数,地理空间实体分为点、线、面、表面和体五类地理空间实体。目前,针对线、面、表面和体地理空间实体开展聚类分析的研究成果较少,而针对点状地理空间实体设计的聚类分析方法仅考虑了实体的位置特征,没有考虑形状特征,不适合面状地理空间实体的聚类,不能满足诸如面状居民地综合类应用的需求。设计面状地理实体的聚类准则需要综合考虑地理实体的距离、几何形状相似性和属性特征。文献[6]给出了面状地理实体形状相似性的定义,本文基于该定义的面状地理实体聚类算法拓展了上述定义,提出了顾及距离和形状相似性因素的面状地理实体相似性准则,并设计了基于新准则的面状地理实体聚类算法。

1　基于线段链形状相似性准则的聚类算法

为了便于讨论,引述文献[6]中线段链相似性定义与相似性准则公式。

定义 1　设 $\delta>0$ 是正常数,$0<e<1$ 是实常数。有形如 $\boldsymbol{Y}=\boldsymbol{AX}+\boldsymbol{B}$ 的变换,$\boldsymbol{A}$ 是 n 阶系数矩阵,$\boldsymbol{Y}$、$\boldsymbol{X}$、$\boldsymbol{B}$ 是 $n\times1$ 阶矩阵,使得对于线段链 $\boldsymbol{X}=(x_1,x_2,\cdots,x_m)$与 $\boldsymbol{Y}=(y_1,y_2,\cdots,y_n)$,$x_i(1\leqslant i\leqslant m)$、$y_j(1\leqslant j\leqslant n)$是线段链 $\boldsymbol{X}$、$\boldsymbol{Y}$ 的顶点,有 $\boldsymbol{X}'=(x_{i_1},\cdots,x_{i_l})$和 $\boldsymbol{Y}'=(y_{j_1},\cdots,y_{j_l})$是 $\boldsymbol{X}$、$\boldsymbol{Y}$ 中满足下述条件的最长子串:① 对 $1\leqslant k\leqslant l-1$,$i_k<i_{k+1}$,$j_k<j_{k+1}$;② 对 $1\leqslant k\leqslant l$,$|i_k-j_k|\leqslant\delta$;③ $1\leqslant k\leqslant l$,$y_{j_k}/(1+e)\leqslant a_ix_{i_k}+b_i\leqslant y_{j_k}(1+e)$,则线段链 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 的相似性 $\mathrm{sim}_{e,\delta}(\boldsymbol{X},\boldsymbol{Y})$定义为$(S_{Lx}+S_{Ly})/(S_X+S_Y)$,$S_{Lx}$、$S_{Ly}$分别表示相似 $\boldsymbol{X}$、$\boldsymbol{Y}$ 线段

收稿日期:2009-01-15。

项目来源:国家 863 计划资助项目(2001AA135080)。

的长度之和；S_X、S_Y 分别表示 $\boldsymbol{X}$、$\boldsymbol{Y}$ 的长度。

对于面状地理实体的聚类问题，线段链顶点 x_i 可用二维坐标表示，其变换矩阵也需要限制才能使相似性定义满足实际应用情况。这里进一步提出满足旋转与平移不变性的线段链相似性评价方法的相似性准则为：

$$|(a_i x_{i_k}+b_i)-y_{j_k}|\leqslant e \qquad (1)$$

1.1 基于线段链形状相似性的面状实体聚类算法

为了验证满足旋转与平移不变性的线段链相似性评价方法的合理性与实用性，以式(1)作为线段链形状相似性准则，设计了基于该准则的面状地理实体聚类算法，称为基于线段链形状相似性准则的聚类算法(clustering algorithm based on a criterion of shape similarity between line segments，CACSS)。算法如下：① 寻找一个面状地理实体 coreobject，要求与该实体几何形状相似的其他面状地理实体数大于2，将 coreobject 设为聚类中心；② 如果 coreobject 存在，依据满足旋转与平移不变性的线段链相似性评价方法，递归搜索与 coreobject 形状相似的面状地理实体；③ 搜索下一个 coreobject；④ 重复步骤②与③。

步骤①和③的计算复杂度在极端情况下(如面状地理实体彼此之间形状均不相似)为 $O(n^2)$，其中，n 是聚类对象个数，步骤②的计算复杂度是 $O(n)$。

利用与文献[5]中相同的实验数据，并去除孤立点，算法 CACSS 聚类得到的结果如图1所示，一共产生4个簇。其中，无线段连接的地理实体不属于任何簇。从图1可见，具有几何形状部分相似性的典型对象聚合到一个簇，如图2所示，黑色表示地理实体是典型的城镇街区，街道两边的居民区的几何形状沿街道部分基本平行，两居民区的几何形状满足旋转与平移不变性的线段链相似性条件，因此被聚合到一个簇中。但从图2也可以发现，两个灰色表示的面状地理实体(椭圆包围)具有相似的几何特征，却没有聚合到该簇中。实验表明，这两个面状地理实体的多边形经旋转、平移后，其彼此面对弧段链的情形如图3所示，有部分弧段链彼此平行，但平行的线段链之间的距离大于参数 e，不能满足线段链相似性准则的要求。

图1 算法 CACSS 聚类结果
Fig. 1 Clustering Result of Algorithm CACSS

图2 图1的局部放大
Fig. 2 Enlarging Effect of Fig. 1

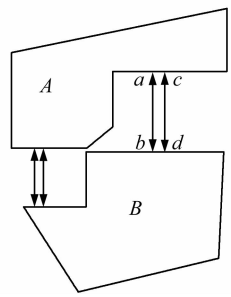


图3 线段链相似性准则的拓展
Fig. 3 Extension of Similarity Criterion Between Line Segments

1.2 线段链相似性准则的拓展

如图3所示，按定义1的要求，面状地理实体 A 与 B 的几何形状不相似。若将式(1)改写为：

$$|(a_i x_{i_k}+b_i)-y_{j_k}|\leqslant e+C(C\text{ 是常数}) \qquad (2)$$

则图3中的面状地理实体 A 与 B 满足该条件。将式(2)作为线段链部分相似性准则，利用算法 CACSS 对§1.1中同样的数据集进行聚类，得到的聚类结果如图4所示。其中，无线段连接的地理实体总是表示它不属于任何簇，属于某个簇的面状地理实体的表示颜色随机产生。

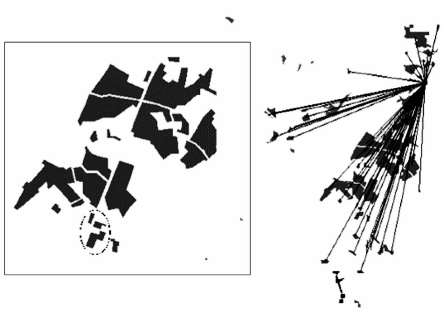


图4 依据线段链相似性准则拓展后的聚类结果
Fig. 4 Clustering Result Based on Extensional Similarity Criterion Between Line Segments

从图4也可以发现，灰色表示的簇所包含的地理实体分布范围跨度大，主要原因是算法 CACSS 并没有考虑距离因素，只有综合考虑了距离和形状相似性得到的聚类结果才是合理的。

2 顾及距离与几何形状相似的面状地理实体聚类

如前所述，面状地理实体聚类时，地理实体相似性需要综合考虑实体之间的距离和几何形状相似性。在此，提出满足该要求的面状地理实体相

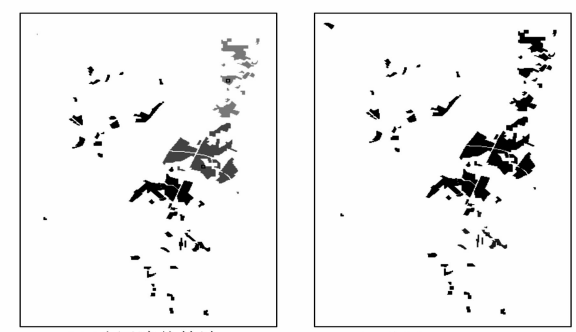
似性准则:

$$\mathrm{sim}_{i,j}=\frac{d_{i,j}}{e^{\mathrm{geomsim}_{i,j}\times C}},0\leqslant i,j\leqslant n \qquad (3)$$

式中,$\mathrm{sim}_{i,j}$表示第 i 个面状地理实体与第 j 个面状地理实体之间的相似性度量;$d_{i,j}$表示第 i 个面状地理实体与第 j 个面状地理实体之间的最近距离;$\mathrm{geomsim}_{i,j}$表示第 i 个面状地理实体与第 j 个面状地理实体之间的几何形状相似性,其值由式(2)和定义 1 计算得到;C 是一常数;n 是参与聚类的面状地理实体个数。

如果实体 i 与实体 j 之间的几何形状相似性为 0,则 $\mathrm{sim}_{i,j}$的值完全由 i 与 j 间的距离来决定;否则,i 与 j 间的几何形状相似性 $\mathrm{geomsim}_{i,j}$的值越大,则 $\mathrm{sim}_{i,j}$值越小,表示 i 与 j 越相似。C 值用来强调或减弱几何形状相似性的作用,$C\geqslant 1$,则强调几何形状相似性的作用;否则,减弱几何形状相似性的作用。

基于式(3)重新设计文献[5]中基于遗传算法的面状地理实体聚类算法,将适应度函数式中的 $\|\cdot\|$ 范数的计算改为用式(3)计算,并以式(3)计算对象之间的相似性替代欧氏距离重写文献[4]中的 CLARANS 算法,分别对文献[5]中相同的实验数据去除孤立点后进行聚类($C=1.0$,聚类数 $k=12$),得到的实验结果如图 5 所示。



(a) 改写遗传算法聚类结果　　(b) 改写 CLARANS 算法聚类结果

图 5　聚类结果

Fig. 5　Clustering Result

从图 5 可见,聚类得到的簇所包含的对象之间的距离比图 4 小得多,彼此距离较大的实体被聚合到不同的簇中;同时,具有形状相似性的对象基本上被聚合到相同的簇中,表明距离和几何形状相似性在聚类时分别"表达"了各自的重要性。

为了验证常数 C 的作用,设 $C=10$,强调几何形状相似性,得到的聚类结果如图 6 所示(聚类数 $k=12$)。从图中可见,右下部分的两个簇中,地理实体基本上是按照几何形状相似性分配到不同的簇中,实体间的距离基本没有起太大作用。设置不同的 C 值,起到了强调或减弱几何形状相似性在聚类时发挥作用大小的效果。

图 6　$C=10$ 时改写遗传算法聚类结果

Fig. 6　Clustering Result Based on Modification of Genetic Algorithm ($C=10$)

3　结　语

通过分析面状地理实体的特征,认为面状地理实体聚类时,地理实体之间的相似性度量仅仅考虑实体之间的距离或仅仅考虑实体之间的几何形状相似性都是不全面的,需要综合考虑这两个因素。依据满足旋转与平移不变性的线段链相似性评价方法设计了面状地理实体聚类算法 CACSS。该算法在不考虑距离因素的前提下,既能得到聚类结果,又能得到聚类数。实验证明,改写后的算法更加适合面状地理实体的聚类分析。

参　考　文　献

[1] Gao X B, Xie W X. Research Progress of Fuzzy Clustering Theory and Application[J]. Chinese Science Bulletin, 1999, 21(44): 2 241-2 251

[2] Hall L O, Ozyurt B, Bezdek J C. Clustering with a Genetically Optimized Approach[J]. IEEE Transactions on Evolutionary Computation, 1999, 3(2): 103-112

[3] Kaufman L, Rousseeuw P J. Finding Groups in Data: An Introduction to Cluster Analysis[M]. New York: Wiley & Sons, 1990

[4] Ng R T, Han J. Efficient and Effective Clustering Methods for Spatial Data Mining[C]. Int Conf Very Large Data Bases (VLDB'94), The University of British Columbia, Vancouver B C, Canada, 1994

[5] 杨春成,张清浦,田向春,等. 基于遗传算法的面状地理实体聚类[J]. 地理与地理信息科学,2004(3):12-16

[6] 杨春成,张清浦,田向春,等. 应用于面状地理实体聚类分析的线段链形状相似性准则[J]. 武汉大学学报·信息科学版,2005,30(1):61-64

第一作者简介:杨春成,研究员,博士。现主要从事空间数据库与数据挖掘方面的研究工作。
E-mail:ycc_3000_vip_ycc@163.com

Clustering Analysis of Geographical Area Entities Considering Distance and Shape Similarity

YANG Chuncheng[1] *HE Liesong*[1] *XIE Peng*[1] *ZHOU Xiaodong*[1]
(1 Xi'an Research Institute of Surveying and Mapping, 1 Middle Yanta Road, Xi'an 710054, China)

Abstract: Geographical area entities are different from geographical point entities, because they have both position feature and shape feature. It is not enough for geographical area entities to be clustered if the clustering criterion just considers distance factor. The clustering criterion designed by us includes distance factor and geometry shape similarity factor. On the basis of this, the corresponding clustering algorithm was implemented. The experimental results show that the algorithm fits to clustering analysis of geographical area entities.

Key words: spatial clustering; geographical area entities; similarity criterion

About the first author: YANG Chuncheng, researcher, Ph.D, majors in spatial database and spatial data mining.
E-mail: ycc_3000_vip_ycc@163.com

(上接第 334 页)

[8] Cover T M, Hart P E. Nearest Neighbor Pattern Classification [J]. Knowledge Based Systems, 1995, 8(6): 373-389
[9] Zhou Shuigeng, Zhao Yue, Guan Jihong, et al. A Neighborhood-based Clustering Algorithm [M]. Berlin/Heidelberg: Springer, 2005
[10] Xiong Hui, Shekhar S, Huang Yan, et al. A Framework for Discovering Co-location Patterns in Data Sets with Extended Spatial Objects[C]. The 4th SIAM International Conference on Data Mining, Florida, USA, 2004

第一作者简介:边馥苓,教授,博士生导师。现从事 GIS 理论及应用研究。
E-mail:wanyou9@gmail.com

A Novel Spatial Co-location Pattern Mining Algorithm Based on *k*-Nearest Feature Relationship

BIAN Fuling[1] *WAN You*[1]
(1 Research Center of Spatial Information and Digital Engineering, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)

Abstract: We define a *k*-nearest feature based on co-location patterns, and develop *k*-nearest feature co-location mining(KNFCOM) algorithm to mine this kind of co-location patterns. The experimental results show that KNFCOM algorithm is efficient and scalable for mining spatial co-location patterns from various large spatial datasets.

Key words: *k*-nearest feature; spatial co-location pattern; KNFCOM; spatial association rule

About the first author: BIAN Fuling, professor, Ph.D supervisor, majors in the theory and application of GIS.
E-mail: wanyou9@gmail.com