文章编号:1671-8860(2011)07-0862-05　　文献标志码:A

空间和属性双重约束下的自组织空间聚类研究

焦利民[1,2]　洪晓峰[1]　刘耀林[1,2]

(1　武汉大学资源与环境科学学院,武汉市珞喻路 129 号,430079)
(2　武汉大学地理信息系统教育部重点实验室,武汉市珞喻路 129 号,430079)

摘　要:形式化定义了双重聚类的聚类准则及其判定方法,提出了双重聚类的两步法求解思路和自组织双重聚类算法。通过实例验证了该算法的可行性,自组织双重聚类可以发现非空间属性的聚集、延伸等空间分布特征,可以发现任意复杂形状的聚类,并降低了人为影响。

关键词:空间聚类;空间和属性约束;双重聚类;自组织网络;属性距离
中图法分类号:P208; P273

带有非空间属性的空间数据聚类分析是空间聚类研究的热点和难点。Lin 等首次使用双重聚类(dual clustering)来指代这样一类空间聚类问题:聚类结果中各子类在空间域上连续、在属性域上相近[1]。由于空间域和属性域的不可比性,对聚类算法中空间距离度量进行属性扩展和对属性距离进行空间扩展都具有一定的人为任意性。双重聚类必须把空间位置变量和属性变量区别对待,由前者派生的空间关系是约束条件,而后者才是聚类目标函数的主要关注域。双重聚类结果可能会产生非凸、环、岛等各种复杂区域,其本质是发现空间连续约束条件下的属性聚集分布特征。

常规空间聚类算法主要有划分法、层次法、基于密度的方法和基于网格的方法等[2-5]。在常规空间聚类算法中扩展非空间属性处理,如 DBRS[4]、GDBSCAN[6]、WaveCluster[7]、CLIQUE[8]、DBSCAN[9]等。对常规空间聚类算法进行非空间属性扩展,或对常规聚类算法进行空间变量扩展,均没有从本质上改变其聚类目标和聚类准则,不能很好地解决双重聚类问题。有学者研究了带约束的空间聚类问题,如数量约束 k-means 聚类[10]、障碍约束条件下的空间聚类[11]等,由于约束条件的不同,这些算法都不能够用来解决双重聚类问题。

李新运等提出了位置距离和属性距离加权的空间距离定义并将其用于聚类问题[12]。Lin 等采用支持向量机(SVM)和层次法属性聚类来实现空间域和属性域上的聚类[1],但是 SVM 的参数对聚类结果影响较大[13]。李光强等提出了基于双重距离和染色法递归检索的空间聚类算法[14],相关研究还有基于惩罚空间距离的聚类[15]、基于密度聚类算法的改进算法等[16]、顾及距离和形状的聚类[17]、多因素模糊聚类[18]等。现有研究提出了聚类结果空间连续、非覆盖(non-overlapping)的条件,但没有给出其判定方法;也没有提出属性内聚性的衡量标准,算法设计上缺少对子类属性内聚性的约束。

1　双重聚类相关概念及其数学定义

双重聚类是同时在空间域和属性域上的聚类[1],要求聚类结果在空间域上连续、属性域上相近。双重聚类旨在发现地理空间对象的属性或属性组合在空间上聚集、延伸、变化的分布规律。双重聚类的相关概念包括空间连续、类内属性距离等。

设有空间点集:

$$F=\{p_1,p_2,\cdots,p_N\},N\geqslant 2 \tag{1}$$

$$p_n=\{g_n^{(x)},g_n^{(y)},a_n^{(1)},\cdots,a_n^{(T)}\},T\geqslant 1 \tag{2}$$

式中,p_n 表示空间点,$g_n^{(x)}$、$g_n^{(y)}$ 表示空间点 p_n 的

收稿日期:2011-04-28。
项目来源:国家自然科学基金资助项目(40901188);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(200906);中央高校基本科研业务费专项基金资助项目(4082002)。

空间位置坐标,$a_n^{(1)},\cdots,a_n^{(T)}$ 表示空间点 p_n 的非空间属性变量。

生成空间点集的 Voronoi 图,以 Voronoi 多边形之间的共边或共点情况作为判断点之间是否空间相邻的标准[19]。以共边作为判断标准,称之为 Rook 准则;以共点作为判断标准,称之为 Queen 准则[20]。本文采用 Rook 准则。

定义 1　空间相邻:若空间点集中的两点 p_i、p_j 的 Voronoi 多边形有公共边,则称 p_i、p_j 空间相邻,记为 $p_i \overset{V}{\longleftrightarrow} p_j$。

定义 2　空间连续:对于空间点子集 F_k 中任意两点 p_i、p_j,至少存在一条彼此空间相邻路径,使得 $p_i \overset{V}{\longleftrightarrow} p_{k_1}, p_{k_1} \overset{V}{\longleftrightarrow} p_{k_2},\cdots, p_{k_t} \overset{V}{\longleftrightarrow} p_j$,则称子集 F_k 是空间连续的。

定义 3　属性距离:空间点集中的两点 p_i、p_j 的非空间属性距离为 $D_{i,j}^{(\mathrm{attr})}=\sqrt{\sum_{t=1}^{T} w_t\,(a_{it}-a_{jt})^2}$,其中 a_{it}、a_{jt} 分别为点 p_i、p_j 的第 t 个属性值,T 为属性个数,w_t 表示第 t 个属性值的权重,$w_1+w_2+\cdots+w_T=1$。

定义 4　类中心:若空间点子集 F_k 构成一个聚类子类,则称虚拟点 $\overline{p^{(F_k)}}(\overline{x^{(F_k)}},\overline{y^{(F_k)}},\overline{a_1^{(F_k)}},\cdots,\overline{a_n^{(F_k)}})$ 为子类 F_k 的类中心,其中 $\overline{x^{(F_k)}}$、$\overline{y^{(F_k)}}$ 分别为子集 F_i 中的横坐标均值和纵坐标均值,$\overline{a_1^{(F_k)}},\cdots,\overline{a_n^{(F_k)}}$ 分别为子集 F_k 中的各属性均值。

定义 5　类内属性距离:若空间点子集 F_k 构成一个聚类子类,则称 $D_{\mathrm{intra}}^{(F_k)}=\frac{1}{N^{(F_k)}}\sum_{n=1}^{N^{(F_k)}} D_{n,\overline{p^{(F_k)}}}^{(\mathrm{attr})}$ 为 F_k 的类内属性距离,其中 $D_{n,\overline{p^{(F_k)}}}^{(\mathrm{attr})}$ 表示点 x_n 和类中心 $\overline{p^{(F_k)}}$ 的属性距离,$N^{(F_k)}$ 为 F_k 内的空间点个数。

基于上述定义,双重聚类的聚类准则可表述为:① 空间连续性,即聚类划分的每一个子类 F_i 均是空间连续的。② 属性内聚性,即每一个子类 F_i 的类内属性距离 $D_{\mathrm{intra}}^{(F_i)}$ 小于给定的阈值 $D_{\mathrm{intra}}^{(\max)}$。

2　自组织双重聚类算法

2.1　双重聚类的求解策略

空间距离和属性距离分别可用于度量聚类对象在空间域和属性域上的邻近关系。但是由于空间域和属性域是性质不同的两个域,无论对空间距离和属性距离进行何种复合运算,都带有人为任意性,也无法实现双重聚类的聚类目标。这里采用两步法:① 将聚类空间分解为若干属性均质的簇。进行属性聚类,并生成点集的 Voronoi 图,点集聚类变为 Voronoi 多边形聚类,将属于同一属性类且空间相邻的 Voronoi 多边形合并(合并后的子集称之为簇),一个属性类对应 1 个或多个空间连续的簇。② 簇合并。循环执行以下操作:计算相邻簇之间的属性距离,选择属性距离最小的两个簇,若合并后类内属性距离小于阈值,则将这两个簇合并。循环执行直至没有新的合并操作产生。该求解策略保证了属性均质性和空间连续性,可以发现复杂形状的聚类,同时簇划分简化了算法过程。

自组织特征映射(self-organizing feature map, SOFM)是一种自组织神经网络,由输入层和输出层组成,两层之间全互连接[21]。采用竞争学习机制实现对输入模式(样本集)的自组织分类,特征相似的点在分类空间中也相邻。SOFM 用于聚类问题,可降低人为影响。本文采用自组织特征映射网络进行属性聚类,实现自组织簇划分。采用递归算法实现簇合并后完成自组织双重聚类。

2.2　自组织双重聚类算法

对于空间点集 F,设空间点个数为 N,点的非空间属性个数为 T,F 中所有点的非空间属性值构成属性向量集 $A=\{a_1,a_2,\cdots,a_N\}$。给定最大的类内属性距离阈值 $D_{\mathrm{intra}}^{(\max)}$。

1) 自组织簇划分

① 构造 SOFM,输入层单元个数为 T,对应于 T 个属性,设输出层单元个数为 $O=\mathrm{int}(D_{\mathrm{intra}}^{(F)}/D_{\mathrm{intra}}^{(\max)})+1$,将所有连接权随机初始化为[0,1]内的值;对原始数据集的坐标和属性进行无量纲化处理。

② 在 F 中随机选取第 n 点的属性向量 a_n 输入。

③ 采用属性距离来衡量输出层中的最佳匹配节点(获胜节点),设获胜节点为节点 c,则 c 应满足:

$$D_{n,c}^{(\mathrm{attr})}=\min D_{n,j}^{(\mathrm{attr})},\ j=1,\cdots,O \tag{3}$$

④ 按以下规则调整权重 $m_j(t)$:

$$m_j(t+1)=\begin{cases} m_j(t)+\mu(t)h(t)[a_n-m_j(t)], & j\in N_c \\ m_j(t), & j\notin N_c \end{cases} \tag{4}$$

式中,$\mu(t)$ 为学习步长;$h(t)$ 为邻域大小;N_c 表示节点 c 的邻域。

⑤ 若 $n<N$,则 $n\leftarrow n+1$,转至步骤②。

⑥ $t\leftarrow t+1$ 转至步骤②,直至网络收敛。

⑦ 生成点集 F 的 Voronoi 图,按聚类结果将

同类相邻 Voronoi 多边形聚合，得到簇划分结果，并将簇进行唯一性编码。

2）递归簇合并

① 扫描每一个簇，找出与之空间相邻且属性距离最小的簇；将扫描结果记为一个 $I\times 3$ 矩阵 $\boldsymbol{C}=\{c_i,c_m,d_{i,m}^{(\mathrm{attr})}\}$，$i=1,2,\cdots,I$，其中 c_i 表示第 i 个簇的标识号，c_m 表示与第 i 个簇空间相邻且属性距离最小的簇的标识号，$d_{i,m}^{(\mathrm{attr})}$ 表示簇 c_i 和 c_m 的属性距离，I 为簇个数。

② 扫描矩阵 $\boldsymbol{C}$，找出最小的属性距离 $\min d_{i,m}^{(\mathrm{attr})}$，对于该距离对应的两个簇 c_i 和 c_m，若由 c_i 和 c_m 构成的子类的类内属性距离 $D_{\mathrm{intra}}^{(F_{c_i,c_1})}<D_{\mathrm{intra}}^{(\max)}$，则合并 c_i 和 c_m，扫描合并簇及其相邻簇，更新矩阵 $\boldsymbol{C}$；若 $D_{\mathrm{intra}}^{(F_{c_i,c_1})}>D_{\mathrm{intra}}^{(\max)}$，转步骤③。

③ 循环执行步骤②，直至没有新的合并操作产生，递归终止。

根据簇合并结果标识点的类别归属，完成聚类划分。

3 实例研究

以武汉市汉口地区的商业用地价格调查样本数据集为例进行实证研究。该数据集是一个包含 6 894 个点的集合，每个点有 3 个基本属性：横坐标 x、纵坐标 y、商业地价 p，样点分布图见图 1。采用双重聚类将地价样本划分为空间连续、地价相近的均质子类，从而发现地价的空间分布规律。

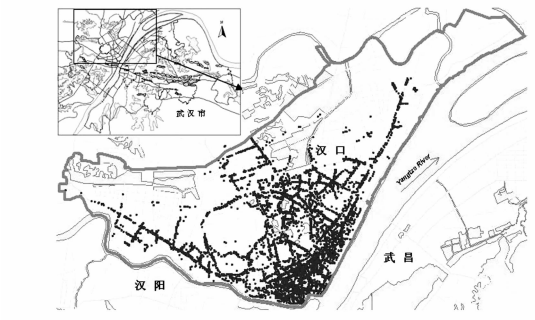

图 1 数据点分布图

Fig. 1 Samples Map

设定聚类的类内属性距离阈值为 800 元/m^2，根据前述方法，自组织簇划分结果如图 2 所示。自组织网络聚类中，输入特征相似的点对应的输出节点也拓扑相邻，因此，图 2 中类别序号变化也反映了簇的属性变化。递归簇合并的结果和聚类结果分别如图 3、图 4 所示。

除采用本文的双重聚类方法，这里还采用基于属性的 k-均值聚类方法、基于位置距离和属性距离加权的复合距离的聚类方法进行比较，聚类结果分别如图 5、图 6 所示。

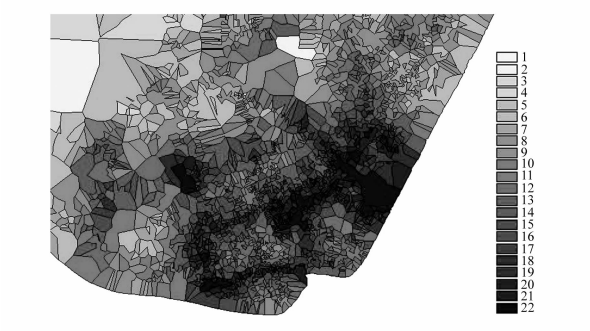

图 2 自组织簇划分

Fig. 2 Self-organizing Sub-clustering

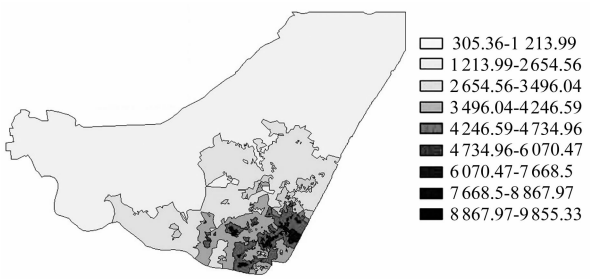

图 3 递归簇合并结果

Fig. 3 Recursive Merging of Sub-clusters

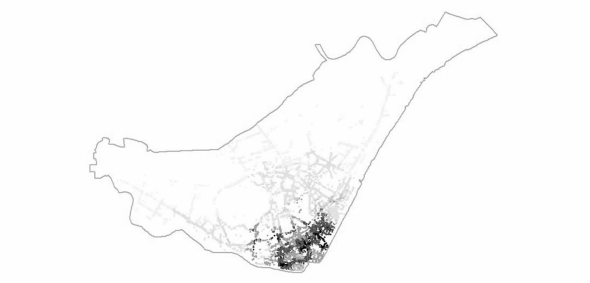

图 4 双重聚类结果

Fig. 4 Result of Dual Clustering

图 5 属性 k-均值聚类结果

Fig. 5 Result of k-means Clusting

对比上述结果发现，属性聚类没有考虑点的空间关系，不能保证聚类的空间连续性，难以清晰展现属性的空间分布规律。双重聚类对数据集的划分实现了空间连续、属性相近，将研究区域划分

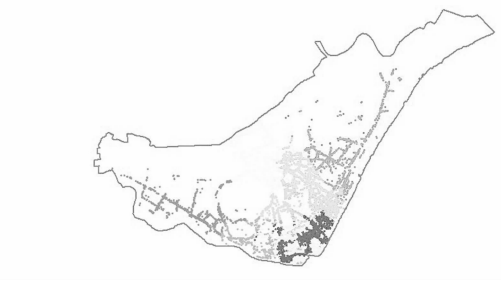

图 6　加权复合距离聚类结果

Fig. 6　Result of Hybrid Distance Based Clustering

成了属性均质区块,发现了属性的空间分布特征。

基于加权距离的聚类结果难以发现复杂形状的聚类,如属性的延伸变化,图 7 中 A、B、C 区域属性相近但由于空间距离较大而被分割成不同类别,同样还有 D、E、F 区域等。基于加权距离的聚类也不能保证子类的属性均质性,不能保证类内属性距离小于某一阈值。双重聚类的簇划分和有属性距离阈值约束的簇合并使得可以发现复杂形状的聚类,并保证类内的属性均质性。类内属性距离阈值体现了用户对于聚类程度的控制,实践中应根据需要和数据情况合理确定,阈值较大则子类较少,阈值较小则子类较多,改变阈值可得到多层次聚类。

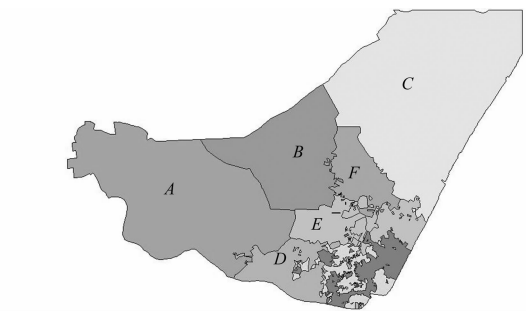

图 7　加权复合距离聚类结果的空间分布

Fig. 7　Distribution of Hybrid Distance Based Clustering Result

4　结　语

双重聚类旨在发现非空间属性的空间聚集、分布、变化规律,体现在聚类准则上就是空间连续性和属性内聚性。双重聚类的目标和准则不同于常规的空间聚类和扩展非空间属性的高维非空间聚类。本文界定了双重聚类的概念,以空间连续、类内属性距离等概念为基础,形式化定义了双重聚类的聚类准则及其判定方法,提出了双重聚类的两步法求解策略和自组织求解算法。聚类实例验证了该算法的可行性,自组织双重聚类可以发现非空间属性的聚集、延伸等空间分布特征,并划分均质区块。该算法降低了人为影响,并可以发现任意复杂形状的聚类。双重聚类可应用于不同领域的高维空间数据挖掘。此外,实践中可能存在异常样本,会使得聚类结果中出现碎小的或岛状的聚类,对异常样本的检测和分析在未来的研究中也应予以重视。

参　考　文　献

[1] Lin C R, Liu K H, Chen M S. Dual Clustering: Integrating Data Clustering over Optimization and Constraint Domains[J]. IEEE Trans Knowledge and Data Engineering, 2005, 17(5):628-637

[2] Han J, Kamber M, Tung A K H. Spatial Clustering Methods in Data Mining: a Survey//Miller H, Han J, eds. Geographic Data Mining and Knowledge Discovery[M]. London:Taylor and Francis, 2001

[3] Halkidi M, Batistakis Y, Vazirgiannis M. On Clustering Validation Techniques[J]. Intelligent Information Systems Journal, 2001, 17(2/3):107-145

[4] Chawla S, Shekhar S, Wu W, et al. Modeling Spatial Dependencies for Mining Geospatial Data: an Introduction//Miller H, Han J, eds. Geographic Data Mining and Knowledge Discovery[M]. London: Taylor and Francis, 2001

[5] Wang X, Hamilton H J. DBRS: A Density Based Spatial Clustering Method with Random Sampling [C]. The 7th PA-KDD, Seoul, Korea, 2003

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

[7] Sheikholeslami G, Chatterjee S, Zhang A. WaveCluster: A Multi-Resolution Clustering Approach for very Large Spatial Databases[C]. The 24th International Conference on very Large Data Bases, New York City, 1998

[8] Agrawal R, Gehrke J, Gunopulos D, et al. Automatic Subspace Clustering of High Dmensional Data for Data Mining Applications[C]. 1998 ACM-SIGMOD Int Conf Management of Data(SIGMOD'98), Settle WA, 1998

[9] 孙志伟,赵政. DBSCAN 在非空间属性处理上的扩展[J],计算机应用,2005,25(6):1 379-1 381

[10] Bradley P S, Bennett K P, Demiriz A. Constrained k-means Clustering, Technical Report MSR-TR-2000-65[R]. Microsoft Research, 2000

[11] Estivill-Castro V, Lee I. Autoclust^{+}: Automatic Clustering of Point-Data Sets in the Presence of Ob-

stacles[C]. Int'l Workshop on Temporal, Spatial and Spatio-Temporal Data Mining, Lyon, France, 2000

[12] 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

[13] Tai C H, Dai B R, Chen M S. Incremental Clustering in Geography and Optimization Spaces[C]. PAKDD'07, Nanjing, China, 2007

[14] 李光强,邓敏,程涛,朱建军. 一种基于双重距离的空间聚类方法[J]. 测绘学报,2008,37(4):482-488

[15] Zhang B, Yin W J, Xie M, et al. Geo-spatial Clustering with Non-spatial Attributes and Geographic Non-overlapping Constraint: a Penalized Spatial Distance Measure[C]. PAKDD'07, Nanjing, China, 2007

[16] Zhou J G, Guan J H, Li P X. DCAD: a Dual Clustering Algorithm for Distributed Spatial Databases [J]. Geo-spatial Information Science, 2007, 10(2): 137-144

[17] 杨春成,何列松,谢鹏,等. 顾及距离与形状相似性的面状地理实体聚类[J]. 武汉大学学报·信息科学版,2009,34(3):335-338

[18] 钟业勋,胡宝清,乔俊军. 多因素评价体系的模糊聚类分析[J]. 武汉大学学报·信息科学版,2010,35(6):752-755

[19] Aurenhammer F. Voronoi Diagrams——a Survey of a Fundamental Geometric Data Structure[J]. ACM Computing Surveys, 1991, 23(3):345-405

[20] Anselin L. Spatial Econometrics: Methods and Models[M]. Dordrecht: Kluwer Academic Publishers, 1988

[21] 张乃尧,闫平凡. 神经网络与模糊控制[M]. 北京:清华大学出版社, 1998

第一作者简介:焦利民,博士,主要研究方向为空间数据挖掘、地理信息的智能化处理。
E-mail:lmjiao027@163.com

Self-organizing Spatial Clustering Under Spatial and Attribute Constraints

JIAO Limin[1,2] *HONG Xiaofeng*[1] *LIU Yaolin*[1,2]

(1 School of Resource and Environmental Science, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)
(2 Key Laboratory of Geographic Information System, Ministry of Education, 129 Luoyu Road, Wuhan University, Wuhan 430079, China)

Abstract: Spatial clustering under spatial and attribute constraints is the clustering analysis on the spatial dataset with non-spatial attributes, which is named dual clustering. The result of dual clustering should be spatially continuous and attributively aggregative. The essence of dual clustering is to find out the clustering and distribution rules of non-spatial attributes. This paper presents the formalized definition of dual clustering, proposes the two-step strategy and self-organizing dual clustering algorithm. Case study verifies the algorithm and shows that the self-organizing dual spatial clustering can find the spatial distribution rules of non-spatial attributes, such as clustering and stretching. Self-organizing dual clustering can detect clusters with complicated shape and reduces the artificial influence.

Key words: spatial clustering; spatial and attribute constraints; dual clustering; self-organizing neural networks; attribute distance

About the first author: JIAO Limin, Ph.D. He is engaged in research on spatial data mining and the intelligent geographic information processing.
E-mail: lmjiao027@163.com