文章编号:1671-8860(2011)08-0982-05　　文献标志码:A

基于力学思想的空间聚类有效性评价

刘启亮[1]　邓　敏[1]　彭东亮[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 932 号,410083)

摘　要:从力学的角度来考虑空间聚类问题,并结合地理学基本规律提出了一种基于力学思想的空间聚类有效性评价指标(简称 SCV)。实验分析表明,本文提出的评价指标能够更准确、高效地对二维地理空间数据的硬聚类结果进行有效性评价。

关键词:空间聚类;有效性评价;力学思想;凝聚力;空间数据挖掘

中图法分类号:P208

空间聚类是一种对空间数据进行深层次分析和挖掘的有效工具[1-4],广泛应用于地理学、制图学、地质学、遥感学、生物学、经济学等众多领域。紧密度和分离度是空间聚类有效性评价的两个基本准则,即空间簇内实体应尽可能紧密,而空间簇间应尽可能分离[5]。现有的聚类有效性评价方法主要分为外部评价法[6]、内部评价法[7]、相对评价法[8,9]。其中,外部评价法和内部评价法均是基于统计学的,计算过程复杂,并且需要过多的先验信息,不适用于空间聚类的有效性评价。相对评价方法不需要借助数据集的先验知识,其主要思想是根据预先定义的评价准则对某种算法不同参数设置的聚类结果进行评价,最终获得最佳的参数组合和聚类模式。在空间聚类时,由于数据集的先验知识一般比较缺乏,同时对参数较为敏感,故相对评价法更适用于评价空间聚类结果的有效性。相对评价法大多采用簇内方差、直径、密度等指标描述簇的紧密程度,采用最短距离、最远距离、重心距离等指标描述簇间的分离程度。相对评价法的有效性评价函数多为紧密度与分离度的组合,即空间簇内部的紧密度越大,与其他簇的分离度越大,聚类质量越高[9]。现有的相对评价方法均难以对包含任意形状空间簇的聚类结果进行有效评价[2,10,11],因此,为了更加有效地对空间聚类结果进行评价,空间聚类有效性评价指标还需满足三方面的要求:① 对空间聚类结果能够进行定量的分析与评价;② 能够评价任意形状空间簇的聚类有效性,同时能够顾及空间异常点的影响;③ 具有较高的运行效率。为此,本文将物理学中的力学思想与地理学规律相结合,发展了一种基于力学思想的空间聚类有效性评价方法。

1　基于力学思想评价空间聚类有效性

现有的空间聚类及其有效性评价方法多是从几何的角度出发,采用各种距离指标度量其紧密度和分离度,缺乏实际的物理意义。本文首先借助 Delaunay 三角网和 Voronoi 图发展了一种适用于空间聚类的力,称之为凝聚力。

1.1　凝聚力

现有的基于引力的聚类评价方法[10]直接根据万有引力来定义实体间的相互作用,并不适用于空间聚类的有效性评价。根据地理学第一定律[12]"越近越相似"的原则,一个空间实体只与其邻近范围内的部分实体有关联,即空间实体间的相互作用力在有限的距离内必须迅速衰减[13]。采用高斯函数定义场强函数[13]虽可以保证力的作用迅速衰减,但由于空间数据分布通常是不均匀的,导致影响因子的设定比较困难。为此,本文

收稿日期:2011-06-15。

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);江苏省资源环境信息工程重点实验室开放研究基金资助项目(JS200901);中南大学前沿研究计划资助项目(2010QYZD002)。

在传统万有引力的基础上,借助 Voronoi 图和 Delaunay 三角网对空间实体邻近关系的自然划分来约束实体间凝聚力的作用范围[14,15]。

二维 Voronoi 图可以视为以平面内的每个点为生长核,以相同速率向外扩张,直到彼此相遇后停止生长,反映了空间实体天然的"势力范围"。本文将 Voronoi 图的这种性质作为对凝聚力的外部约束,即一个空间实体只与其直接 Voronoi 邻近实体之间具有凝聚力作用,这种作用符合距离平方反比关系,与万有引力类似,而与其他实体间的凝聚力作用忽略不计。对于任一空间实体 P,其所有直接 Voronoi 邻近实体集合表示为 NV(P),进而可以表达两个空间实体间的凝聚力为:

$$F_{\text{agg}}(P,Q)=\begin{cases}k\dfrac{m_P m_Q}{d\ (P,Q)^2}, Q\in \text{NV}(P)\\ 0, Q\notin \text{NV}(P)\end{cases} \tag{1}$$

式中,k 为引力系数,可以设为 1;m_P、m_Q 为实体 P、Q 的质量,考虑到可以将空间点实体均视为单位质点,故均令 m_P、m_Q 为 1;$d(P,Q)$ 为实体 P 与 Q 的欧氏距离。

由于 Delaunay 三角网是 Voronoi 图的对偶图,故实际操作中只需采用 Delaunay 三角网定义的空间邻近关系来计算凝聚力,避免了计算的复杂性。在此基础上,进一步发展空间聚类有效性评价方法。

1.2　基于凝聚力的空间聚类有效性评价方法

从凝聚力的角度,如果某个空间簇内的实体间的凝聚力的作用越强,其"抱团"的趋势也就越明显,作为一个空间簇存在也就越合理。而不同簇之间加以区分,是因为簇之间实体的相互凝聚力过小。空间异常点可以视为一个特殊的空间簇,故异常点受到的凝聚力也应该较小。因此,一个高质量的空间聚类结果必须满足以下两个条件:① 空间簇(包括空间异常点)之间凝聚力的作用尽可能小;② 空间簇内实体间的凝聚力作用尽可能大。下面给出有效性评价指标的具体定义。

定义 1　凝聚树:对于空间簇 C,其形成过程可以描述为从任一空间实体开始,不断"吸附"与当前实体凝聚力作用最大的实体,最终簇中的所有实体形成一个通过凝聚力联系起来的树形图结构,称为凝聚树。

凝聚树与最小生成树具有相同的几何结构,凝聚树中所有实体间凝聚力的作用最强,可以更好地反映空间簇实体的紧密程度,进而给出簇内凝聚力的定义。

定义 2　簇内凝聚力:对于一个包含 n 个实体的空间簇 C,其生成的凝聚树为 AT(C),凝聚树中实体间的所有凝聚力构成凝聚力集合 AF(C),凝聚力平均值记为 μ_{AF},方差记为 σ_{AF}。进而,一个空间簇 $C(P_i\in C,\ P_j\in C)$ 的簇内凝聚力 $F_I(C)$ 可以表达为:

$$F_I(C)=\begin{cases}\dfrac{\sum F_{\text{agg}}(p_i,p_j)}{m}, \exists\, F_{\text{agg}}(p_i,p_j)\leqslant \mu_{\text{AF}}-\sigma_{\text{AF}}\\ \mu_{\text{AF}}, \forall\, F_{\text{agg}}(p_i,p_j)>\mu_{\text{AF}}-\sigma_{\text{AF}}\end{cases} \tag{2}$$

式中,m 表示所有小于 $\mu_{\text{AF}}-\sigma_{\text{AF}}$ 的凝聚力数量。从式(5)可看出,簇内凝聚力与空间实体的分布密切相关,本文采用统计量 $\mu_{\text{AF}}-\sigma_{\text{AF}}$ 度量实体的分布情况。当实体分布较均匀时,簇内实体间凝聚力的变化较小,直接以其平均值度量簇的紧密程度;而当实体分布变化较大时,则采用簇中实体间的部分较小凝聚力的平均值度量簇的紧密程度。

定义 3　簇间凝聚力:对于任一空间簇(或空间异常点)C,其簇间凝聚力定义为其他簇内实体对 C 内实体凝聚力的平均值,记为 $F_E(C)$,表达为:

$$F_E(C)=\sum F_C(p_i,p_j)/w,\quad p_i\in C,\ p_j\notin C \text{ 且 } p_j\in \text{NV}(p_i) \tag{3}$$

式中,w 表示其他空间簇(包括空间异常点)与空间簇 C 中实体有凝聚力作用的实体数量。

进而,对于一个空间数据库 SDB,不妨设其被划分为 M 个空间簇,分别记为 C_1,C_2,…,C_M,同时获得 N 个空间异常点(如 DBSCAN[16] 等算法识别的空间孤立点),分别记为 O_1,O_2,…,O_N,于是,空间聚类有效性度量指标可以定义为所有簇内凝聚力的最小值与簇间凝聚力平均值的比值,记为 SCV,表达为:

$$\text{SCV}=(M+N)\min_{i=1,\cdots,M}F_I(C_i)/\left(\sum_{i=1}^{M}F_E(C_i)+\sum_{i=1}^{N}F_E(O_i)\right) \tag{4}$$

从式(4)可以看出,针对一个空间聚类的结果,若空间簇内实体的凝聚力越强,簇间实体的凝聚力作用越弱,异常点受到的凝聚力作用也较弱,则 SCV 值越大,表示聚类的效果越好。同时也可以发现,本文在评价过程中顾及了空间异常点的影响,而且在计算簇间的凝聚力作用时,将空间簇视为一个整体,充分体现了不同簇之间的凝聚力作用关系,避免了传统评价簇间的分离度指标(如最短距离、质心距离)不能适用于任意形状空间簇

的缺点。

1.3 方法实现与复杂度分析

对于一个包含 n 个空间实体的空间数据库SDB,采用SCV指标对其空间聚类有效性进行评价时,主要分为以下几个步骤:① 针对空间数据库中所有实体构建Delaunay三角网,复杂度为 $O(n\lg n)$。② 针对每个空间簇生成凝聚树,进而计算簇内凝聚力。本文采用Kruskal算法[17]在Delaunay三角网的基础上针对每个空间簇分别构建凝聚树,其复杂度为 $O(n\lg n)$;计算簇内凝聚力在最差情况下的复杂度为 $O(n\lg n)$。③ 计算簇间凝聚力。由于Delaunay三角网的边数最大为 $3n-6$,每个结点直接连接的结点数平均为6,于是此步骤的复杂度为 $nO(6\lg 6)=O(n)$。④ 计算有效性评价指标SCV。

综合以上几个步骤,采用SCV指标对空间聚类的有效性进行评价的时间复杂度为 $O(n\lg n)+O(n\lg n)+O(n\lg n)+O(n)\approx n\lg(n)$,而当前聚类有效性指标[10,11]的复杂度多为 $O(n^2)$,因此本文方法的效率明显提高。

2 实验分析

下面设计两个实验来验证本文空间聚类有效性评价方法的正确性和可行性。实验一模拟了文献[17]中采用的三组经典模拟数据库作为研究对象,如图1(a)~1(c)所示,并分别对经典的DBSCAN算法[17]和 K-means算法[1]的空间聚类结果进行有效性评价,指导选择最佳聚类算法和聚类参数组合,其中数据库3包含异常点,用以检验有效性评价指标在包含异常点情况下的评价结果。实验二采用文献[18]整理得到的2008年位于我国东部沿海、中部(河南省)县域经济实力位居全国县级单位前100位的93个县级城市空间分布数据作为研究实例,指导DBSCAN算法选择最佳的聚类参数。图2和图3分别给出了DBSCAN和 K-means算法对三组模拟数据库的聚类结果。

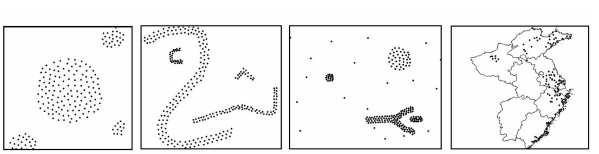


图1 实验数据库

Fig. 1 Experimental Databases

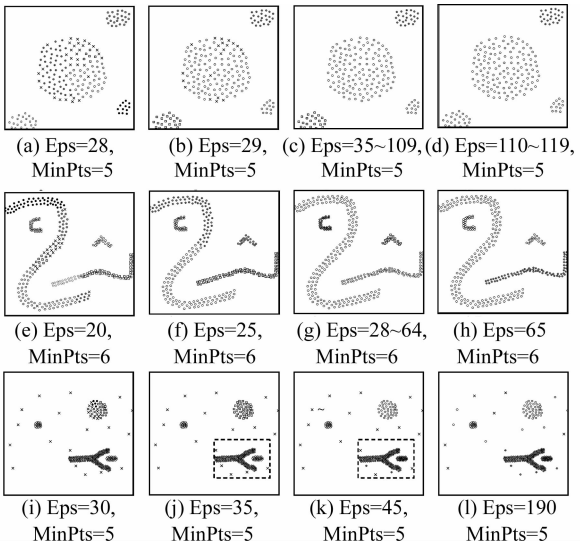


图2 DBSCAN算法聚类结果(×-异常点)

Fig. 2 Results via DBSCAN Spatial Clustering (×-Outlier)

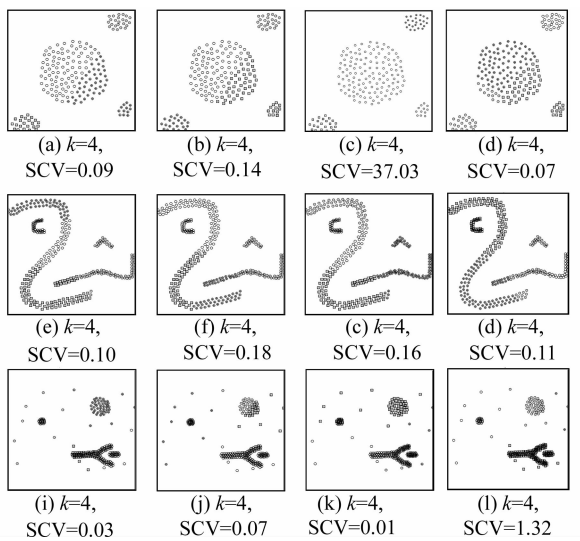


图3 K-means算法聚类结果

Fig. 3 Results via K-means Spatial Clustering

分别采用本文提出的SCV指标和Dunn指数对上述聚类结果进行评价,结果列于表1。Dunn指数采用簇的直径表示空间簇的紧密程度,采用簇间距离表示空间簇间的分离程度,具体表达如下:

$$D_k=\min_{i=1,\cdots,k}\{\min_{j=i+1,\cdots,k}[\frac{d(C_i,C_j)}{\max_{m=1,\cdots,k}(\mathrm{diam}(C_m))}]\}$$

$$d(C_i,C_j)=\min_{x\in C_i,y\in C_j}\{d(x,y)\}$$

$$\mathrm{diam}(C_m)=\max_{x,y\in C_m}\{d(x,y)\} \quad (5)$$

式中,$d(C_i,C_j)$ 为簇间实体的最短距离;$\mathrm{diam}(C_m)$ 为簇的直径,即簇内实体间的最大距离;m 表示聚类簇数。Dunn指标数值越大,表示聚类结果越好。

表 1　DBSCAN 和 *K*-means 聚类的有效性评价结果

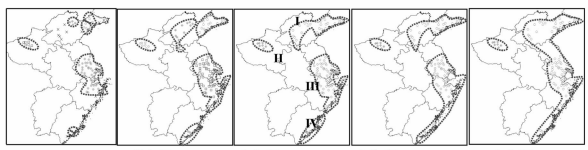
Tab. 1　Validity Assessment of Spatial Clustering Results of DBSCAN and *K*-means

	DBSCAN				*K*-means	
	Eps	MinPts	SCV	Dunn	SCV	Dunn
数据库 1	28	5	1.35	0.10	0.09	0.06
	29	5	1.66	0.35	0.14	0.07
	35～109	5	37.03	0.35	37.03	0.35
	110～119	5	2.32	0.26	0.07	0.07
数据库 2	20	6	2.21	0.05	0.10	0.03
	25	6	15.73	0.05	0.18	0.05
	28～64	6	28.68	0.12	0.16	0.04
	65	6	2.53	0.13	0.11	0.04
数据库 3	30	5	2.28	0.11	0.03	0.03
	35	5	9.63	0.11	0.07	0.06
	45	5	15.21	0.69	0.01	0.03
	190	5	2.69	0.39	1.32	0.31

对 DBSCAN 算法聚类评价结果进行分析可以发现:① 针对数据库 1 和 2,SCV 指标可以准确识别最佳聚类结果,当出现过多异常点或过度聚类时,SCV 指标较低,而在正确聚类时,SCV 指标显著增大,因此可以有效指导聚类参数的选择。而 Dunn 指数对数据库 1、2 的评价结果差异不大,同时无法准确反映聚类的有效性。如对于数据库 1,无法区分图 2(b)和 2(c)的聚类结果;对于数据库 2,则给出了错误的评价结果。② 针对数据库 3, Dunn 指数对图 2(i)、2(j)的聚类结果无法区分,过度聚类时,有效性反而较高,显然不合理;而针对图 2(i)的情况,SCV 指标在评价时顾及了异常点的影响,当空间簇被误判为异常点时,SCV 值较小,可以与正确聚类结果进行区分。根据 SCV 指标可以看出,图 2(j)、2(k)的有效性较高,其中图 2(k)的有效性最高,似乎与文献[17]中的正确聚类结果存在差异,而实际上这是由于空间聚类中存在的尺度效应造成的。在较大尺度上(图 2(k)),虚线内实体的内部差异性不显著,而随着尺度缩小(图 2(j)),则应进行区分。对图 2(j)、2(k)中虚线框部分的聚类结果进行局部评价发现,图 2(j)中聚类结果的有效性(SCV=14.25)明显大于图 2(k)的情况(SCV=6.69)。因此,上述分析可以充分说明,SCV 指标可以更加有效地指导识别最佳的聚类结果。

进一步分析 DBSCAN 和 *K*-means 算法对同一数据库的聚类结果可以发现,*K*-means 算法的聚类有效性远低于 DBSCAN 算法。在本文实验中,只在数据库 1 中可以得到较好的聚类效果,而对于任意形状的空间簇以及包含较多空间异常点的情况,效果很不理想,这也表明上述数据库选用的 DBSCAN 算法聚类更为合理,该结论与当前研究中的论断[1,17]也是相符的。

下面采用实际地理空间数据对 SCV 指标的实用性进行验证。如图 1(d)所示的县级城市,其经济发展水平明显高于全国平均水平,现采用 DBSCAN 算法对其空间聚集模式进行挖掘,不同参数下 DBSCAN 算法的聚类结果如图 4 所示,空间簇采用虚线标出,SCV 指标与 Dunn 指数的有效性评价结果列于表 2。



(a) Eps=100 km MinPts=3　(b) Eps=200 km MinPts=3　(c) Eps=300 km MinPts=3　(d) Eps=500 km MinPts=3　(e) Eps=600 km MinPts=3

图 4　DBSCAN 算法聚类结果(×-异常点)

Fig. 4　Results via DBSCAN spatial Clustering(×-Outlier)

表 2　利用 SCV 指标和 Dunn 指数评价空间聚类有效性

Tab. 2　Validity Assessment of Spatial Clustering Results by SCV and Dunn Indicators

Eps/km	MinPts	SCV	Dunn
100	3	2.63	0.45
200	3	3.37	0.36
300	3	17.36	0.39
500	3	3.61	0.30
600	3	3.60	0.24

根据 SCV 指标的评价结果可以发现,在当前尺度上 Eps=300 km、MinPts=3 时的聚类结果最佳,即图 4(c)所示的聚类结果,对此时获得的空间簇结构进行分析可以发现,聚类结果比较合理,大致可以分为山东环渤海城市群(空间簇 I)、中原城市群(空间簇 II)、江浙城市群(空间簇 III)以及华南沿海城市圈(空间簇 IV),与我国城市发展的空间分布[18]非常吻合。而根据 Dunn 指数的评价结果,当 Eps=100 km、MinPts=3 时,聚类结果最优,即图 4(a)所示的聚类结果,此时的聚类结果误判了过多的空间异常点,显然是不合理的。由此可以说明,SCV 指标对于实际应用中指导聚类最佳参数的选择更加有效和实用。

3　结　语

空间聚类有效性评价对于实际应用中选择最佳的聚类算法以及选择最优的聚类参数具有重要的意义。本文提出了一种基于力学思想的空间聚类有效性评价方法。通过实验分析以及与经典的

Dunn指数的比较可以发现,本文方法具有以下优点:① 可以对任意形状空间簇的聚类有效性进行准确评价,同时很好地顾及了空间异常点对聚类结果的影响;② 可以有效指导最佳聚类算法和聚类参数的选择;③ 效率较高,时间复杂度仅为$n\lg(n)$。

有待进一步研究的工作主要集中在以下两个方面:① 本文方法能够针对空间分布变化不大的聚类结果进行有效评价,进一步需要研究其针对空间分布差异较大的空间聚类结果的评价效果。② 需要发展一种同时顾及空间属性与专题属性相似性的一体化空间聚类度量方法,这也是当前空间聚类与聚类有效性评价研究中的一大难题,本文提出的有效性度量策略将为解决该问题提供重要的理论与方法基础。

参 考 文 献

[1] Miller H, Han J. Geographic Data Mining and Knowledge Discovery[M]. 2nd ed. Boca Raton: CRC Press, 2009

[2] 李晓雯,毛政元,李建微. 一种基于几何概率的聚类有效性函数[J]. 中国图像图形学报, 2008, 13(12): 2 351-2 356

[3] 李德仁,王树良,李德毅,等. 论空间数据挖掘和知识发现的理论和方法[J]. 武汉大学学报·信息科学版, 2002, 27(3): 221-233

[4] 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

[5] Berry M, Linoff G. Data Mining Techniques for Marketing, Sales and Customer Support[M]. New York: John Wiley & Sons Inc, 1996

[6] Fowlkes E, Mallows C. A Method for Comparing Two Hierarchical Clusterings[J]. Journal of the American Statistical Association, 1983, 382(78): 569-576

[7] HalkidiI M, Batistakis Y, Vazirgiannis M. On Clustering Validation Techniques[J]. Intelligent Information Systems, 2001, 223(17): 107-145

[8] Pal N G, Biswas J. Cluster Validation Using Graph Theoretic Concepts[J]. Pattern Recognition, 1997, 30(6): 847-857

[9] Kovacs F, Legany C, Babos A. Cluster Validity Measurement Techniques[C]. The 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, Madrid, Spain, 2006

[10] 于勇前,赵相国,陈衡岳,等. 基于引力概念的聚类质量评估算法[J]. 东北大学学报(自然科学版), 2007, 28(8): 1 109-1 112

[11] 岳士弘,李平,于剑. 一组新的聚类有效性指标[J]. 模式识别与人工智能, 2004, 17(4): 516-522

[12] Tobler W. A Computer Movie Simulating Urban Growth in the Detroit Region[J]. Economic Geography, 1970, 46(2): 234-240

[13] 淦文燕,李德毅,王建民. 一种基于数据场的层次聚类算法[J]. 电子学报, 2006, 34(2): 258-262

[14] 闫超德,赵仁亮,陈军,等. Voronoi图的首最邻近递归收敛特性及其应用[J]. 武汉大学学报·信息科学版, 2008, 33(11): 1 194-1 197

[15] 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): 694-709

[16] 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, OR, 1996

[17] 卜月华,吴建专,顾国华,等. 图论及其应用[M]. 南京:东南大学出版社, 2002

[18] 刘福刚,孟宪刚. 中国县域经济年鉴(2008卷)[M]. 北京:社会科学文献出版社, 2008

第一作者简介:刘启亮,硕士生,研究方向为时空聚类、时空异常探测及其在全球气候变化中的应用。

E-mail:liuqiliang192@126.com

Validity Assessment of Spatial Clustering Methods Based on Gravitational Theory

LIU Qiliang[1] *DENG Min*[1] *PENG Dongliang*[1] *WANG Jiaqiu*[1]

(1 Department of Surveying and Geo-informatics, Central South University, 932 South Lushan Road, Changsha 410083, China)

Abstract: To overcome such limitations, the gravitational theory is firstly employed to describe the issue of spatial clustering, where a force for spatial clustering, called aggregation

(下转第990页)