DOI:10.13203/j.whugis20180043 文章编号:1671-8860(2019)08-1131-07

顾及密度对比的多层次聚类点群选取方法

程绵绵[1] 孙 群[1] 李少梅[1] 徐 立[1]

1 信息工程大学地理空间信息学院,河南 郑州,450001

摘 要:在语义信息缺乏的情况下进行点群选取是制图综合的难点之一。提出了一种新的通过多层次聚类进行点群选取的方法。首先,针对*k*-means聚类算法的不足,利用改进的密度峰值聚类算法实现点群自动聚类,主要表现为用基尼系数确定最优截断距离及用局部密度和相对距离的关系自动确定聚类中心。其次,提出一种顾及密度对比的选取策略,通过点群多层次聚类,将点群划分成不同等级的簇,确定不同等级的聚类中心,建立点群的层次树结构;依据方根定律计算的选取数量,按照各级别簇的点数比例,自上而下逐层分配待选取点数,确定选取对象,实现点群的自动选取和多尺度表达。对不同分布模式的点群进行实验,验证了该方法的普适性和有效性。

关键词:空间聚类;密度峰值;基尼系数;点群选取;制图综合

中图分类号:P208 **文献标志码**:A

点群选取是制图综合的重要内容,目的是在点数减少的情况下,尽可能正确地表达点群的空间信息;难点是在语义信息缺乏时,如何仅仅依赖点群的几何信息进行选取。对此,许多学者进行了研究[1-5],基本思想大都是依据几何信息构造一定的模型,对点的重要性进行判断,从而对点群进行取舍。在这些方法中,点群的聚类及Voronoi图是综合过程的重要手段,在制图综合中具有广泛的应用[6-7]。然而现有方法在对聚类算法本身进行深入考虑方面较为欠缺。如文献[4]通过*k*-means算法提取聚类中心,并用层次Voronoi图逐步细化地表达聚类中心点,实现点群的综合,为点群的多尺度表达提供了一种有效的思路,但存在以下两个方面的问题:一是*k*-means算法需要事先确定*k*值及人为指定初始聚类中心,且*k*-means算法将每个点指派到距离最近的聚类中心,因而不能检测非球面分布的点群数据;二是单纯的聚类只考虑数据本身的分布特点,若用同一层级的聚类中心对点群进行粗略表达,不能顾及点群选取必须满足的密度对比要求。因此,该方法不能实现点群的全自动选取,选取结果也存在一定的不确定性和不合理性。

针对上述不足,本文选用更加合理的聚类算法实现点群的自动聚类,进而提出层次聚类的概念,在此基础上提出一种顾及密度对比的点群选取方法。

1 基于改进密度峰值聚类算法的点群自动聚类

聚类是根据数据对象及其关系对数据进行分组,从而分析数据的潜在结构,判断数据的自然簇属性并压缩数据存储容量的数据分析方法[8-10]。现有的聚类方法主要分为基于划分的聚类[11]、基于层次的聚类[12]、基于密度的聚类[9,13]、基于网格的聚类[14]及基于模型的聚类[15]等,其中,基于密度的聚类方法以高密度区域作为判断依据。这种非参数方法与传统方法相比,不仅适用于处理任何形状的数据集,而且无需预先设定簇的数量,更易于聚类的自动实现[9]。因此,本文基于Rodriguez等[16]提出的一种密度峰值聚类算法,实现空间点群的自动聚类。

1.1 密度峰值聚类算法及分析

密度峰值聚类算法原理是基于聚类中心的两个特点提出的,一是聚类中心本身密度大;二是聚类中心与其他密度更大的点之间的距离相

收稿日期:2018-05-04

项目资助:国家自然科学基金(41571399)。

第一作者:程绵绵,博士生,主要从事多源数据融合及制图综合研究。chmmian@163.com

通讯作者:孙群,博士,教授。sunqun@371.net

对更大。其主要过程是，对于数据集$D=\{p_1,p_2\cdots p_n\}$中的每个点p_i，计算局部密度ρ_i和相对距离δ_i，其中ρ_i表示数据集中到p_i的距离小于或等于d_c的个数，d_c称为截断距离(cutoff distance)，需要用户自行确定。文献[16]利用cut-off核(cut-off kernel)计算ρ_i：

$$\rho_i=\sum_{j\in I_S\backslash\{i\}}\chi(d_{ij}-d_c) \tag{1}$$

式中，d_{ij}表示点p_i与p_j间的欧氏距离；$I_S=\{1,2\cdots n\}$(下同)，$j\in I_S\backslash\{i\}$表示$j\in I_S$且$j\neq i$；函数$\chi(x)$计算方法为：

$$\chi(x)=\begin{cases}1,x<0\\0,x\geqslant 0\end{cases} \tag{2}$$

cut-off核计算的局部密度为离散值，有学者提出利用高斯核(Gaussian kernel)计算局部密度[17]，得到连续值，减小聚类产生冲突的概率，计算公式为：

$$\rho_i=\sum_{j\in I_S\backslash\{i\}}\mathrm{e}^{-(d_{ij}/d_c)^2} \tag{3}$$

相对距离δ_i的计算公式为：

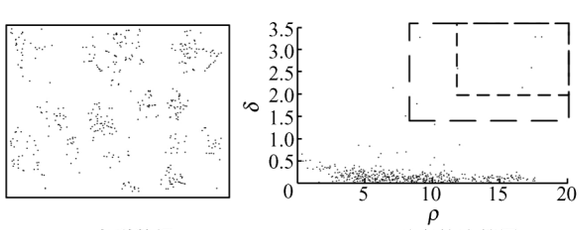
$$\delta_i=\begin{cases}\min\limits_{j\in I_S^i}\{d_{ij}\},I_S^i\neq\varnothing\\\max\limits_{j\in I_S^i}\{d_{ij}\},I_S^i=\varnothing\end{cases} \tag{4}$$

式中，I_S^i为指标集，

$$I_S^i=\{k\in I_S:\rho_k>\rho_i\} \tag{5}$$

计算出每个点的ρ_i和δ_i后，以二元对(ρ_i,δ_i)的形式绘制于二维坐标系中，称之为决策图，如图1所示。从决策图中选取横纵坐标都比较大的点作为密度峰值点，即聚类中心。根据聚类中心及密度边界阈值，将剩余点分到各簇中，完成聚类过程。

密度峰值聚类算法虽有诸多优点，但将该方法应用于空间点群的聚类还存在两个缺点。一是需要凭经验选取截断距离，降低了算法的科学性和普适性；二是需要人工根据决策图确定聚类中心，对于同样的决策图，不同的人可能得出不同结果。



(a) 点群数据 (b) 对应的决策图

图1 点群数据的决策图

Fig.1 Decision Graph of Point Group

如图1(b)中，短虚线框和长虚线框分别选取了不同数量的密度峰值点，则会得到完全不同的聚类结果。

1.2 密度峰值聚类算法的改进

1.2.1 截断距离自适应确定

从式(1)、式(3)可以看出，ρ_i由d_c决定，在文献[16]中，d_c是根据若干数据集的经验值设定的，其选取的标准是使每个数据点的平均邻居个数约为数据点总数的2%。但是对于空间数据，不同数据集在点数及分布模式上都存在差异，因此更科学的方法是根据数据集本身的特点自适应地设置d_c的大小。

文献[18]指出数据的势能与数据的不纯度之间存在一定的关系，即数据的势能分布较均匀时，数据的不纯度较大；数据的势能分布不均匀时，数据的不纯度较小。数据的不纯度最小时，数据的势能差别最大，更易于聚类。对于空间点群集合$P=\{p_i\}_{i=1}^n$，每个点的势能计算[18]公式为：

$$\varepsilon(x)=\sum_{i=1}^{n}\mathrm{e}^{-(d_{xi}/\sigma)^2} \tag{6}$$

式中，σ为影响因子。

数据的不纯度可以由基尼指数计算[18]：

$$G=1-\sum_{i=1}^{n}\left(\varepsilon_i/Z\right)^2 \tag{7}$$

式中，$Z=\sum_{i=1}^{n}\varepsilon_i$，为数据域的总势能；显然$G$是$\sigma$的函数，即$G=G(\sigma)$。

对比点的势能计算式(6)和局部密度计算式(3)发现二者完全等价。因此，为了计算合理的局部密度，最优的截断距离d_c等价于取得最优的影响因子σ。

根据上述分析，以基尼指数值最小时的影响因子作为截断距离的值，能够达到最优聚类的效果，即：

$$d_c=\arg(\min G(\sigma)) \tag{8}$$

1.2.2 聚类中心的自动确定

为了形象说明聚类中心的确定方法，考虑图2(a)中的例子，其中包含32个离散点，将二元对$\{(\rho_i,\delta_i)\}_{i=1}^{32}$在二维直角坐标系中画出，如图2(b)所示。

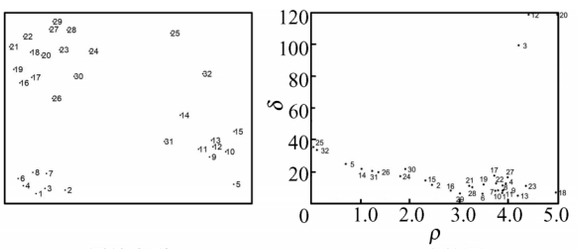
从图2(b)中可以看出，点20、12、3由于同时具有较大的局部密度ρ和相对距离δ，位于坐标系的右上方，并从其他数据点中脱颖而出，图2(a)中这3个点正好可能是原始点群的3个聚类中心。

为了自动探测聚类中心，用ρ和δ归一化的

乘积评估点间的差异度,根据差异度的统计特征和变化规律确定聚类中心。定义簇中心权值为:

$$\gamma_i = \frac{\rho_i - \rho_{\min}}{\rho_{\max} - \rho_{\min}} \cdot \frac{\delta_i - \delta_{\min}}{\delta_{\max} - \delta_{\min}}, i \in I_S \quad (9)$$

式中,$\rho_{\min}$、$\rho_{\max}$、$\delta_{\min}$、$\delta_{\max}$分别是ρ_i、δ_i的最小值、最大值。



(a) 原始点群　　(b) 决策图

图2　原始点群与决策图

Fig.2　Original Point Group and Decision Graph

将γ_i从大到小降序排列,以序号N为横坐标,γ_i为纵坐标,将排序结果在二维直角坐标系中画出,如图3所示。从图3中可以看出,γ总体呈下降趋势,但是下降程度各阶段有所不同:聚类中心的γ值下降比较剧烈,非聚类中心的γ值下降比较平滑,而从聚类中心到非聚类中心过渡时γ值有一个明显的跳跃。为了确定曲线下降的拐点,用两点连线的斜率表示簇中心权值的下降趋势,即:

$$k_i^m = (\delta_{i+m} - \delta_i)/m \quad (10)$$

式中,k_i^m表示在区间$[i, i+m]$上簇中心权值的平均变化率,该参数描述了某一区间内γ的总体变化趋势,则γ值变化的拐点满足:

$$x = \arg(\max(k_1^{i-1}/k_i^{n-i})) \quad (11)$$

根据式(10),k_1^{i-1}表示第1个点到第i个点的斜率,k_i^{n-i}表示第i个点到n个点的斜率。因此,拐点可以认为是簇中心权值总体变化最快的临界点。在二维坐标系中绘出坐标对$(i, k_1^{i-1}/k_i^{n-i})$,如图4所示。显而易见,序号为4的点为拐点,对应图3中点号为17的点为拐点,从而自动探测出聚类中心点为点20、12、3。

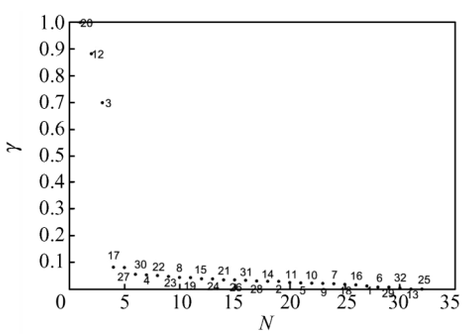


图3　γ排序图

Fig.3　Sorting Chart of γ

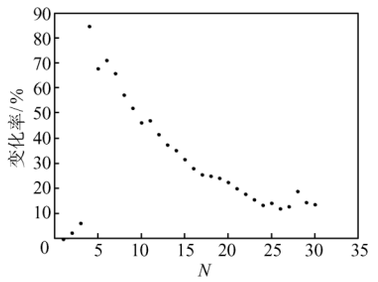


图4　簇中心权值变化率

Fig.4　Change Rate of Cluster Center Weight

2　顾及密度对比的多层次聚类点群综合

2.1　点群多层次聚类

定义1　多层次聚类。对点群$P=\{p_i\}_{i=1}^n$的聚类结果$P=\{\text{cluster}(c_1), \text{cluster}(c_2)\cdots\text{cluster}(c_k)\}$分别进行再次聚类,其任意一个点簇$\text{cluster}(c_j)(j=1,2\cdots k)$的聚类结果为$\text{cluster}(c_j)=\{\text{cluster}(c_j^1), \text{cluster}(c_j^2) \cdots \text{cluster}(c_j^{k'})\}$,其中$\text{cluster}(c_j^l)(j=1,2\cdots k, l=1,2\cdots k')$表示点簇,$c_j^l$为各自聚类中心。重复上述步骤,直到所有簇为单个点的过程,称为点群的多层次聚类。从定义1可以看出,多层次聚类是将上一层级的一个簇作为下一层级的原始聚类点群,进行重复迭代聚类,目的是将点群划分成不同等级的簇,并确定不同等级的聚类中心。点群的多层次聚类结果可以存储为一种层次树结构,且由于点群分布的不均匀性,这种层次树结构表现为非平衡多叉树。将上述算例中的数据进行多层次迭代聚类,得出如图5所示的层次树结构,其中灰色背景标注的点为各级别簇的聚类中心。

2.2　顾及密度对比的点群选取方法

为了使点群选取满足密度对比要求,本文采用以下选取策略。一是根据各级别簇中点数从上而下分配选取的点数循环,直到分配到各簇的选取数量为1;二是只选取聚类中心。通过上述两条策略,能够根据选取点数定位到选取的点。具体步骤如下:

1)对点群进行多层次迭代聚类,生成点群层次树;

2)根据制图综合需要,依据开方根定律计算需要选取的数量N_{select};

3)在第1次聚类的结果上,按照各簇点数比例分配需要选取的数量N_{select};

4)以此类推,逐层往下分配需要选取的数量,直到需要选取的数量均分配到聚类中心点为止。

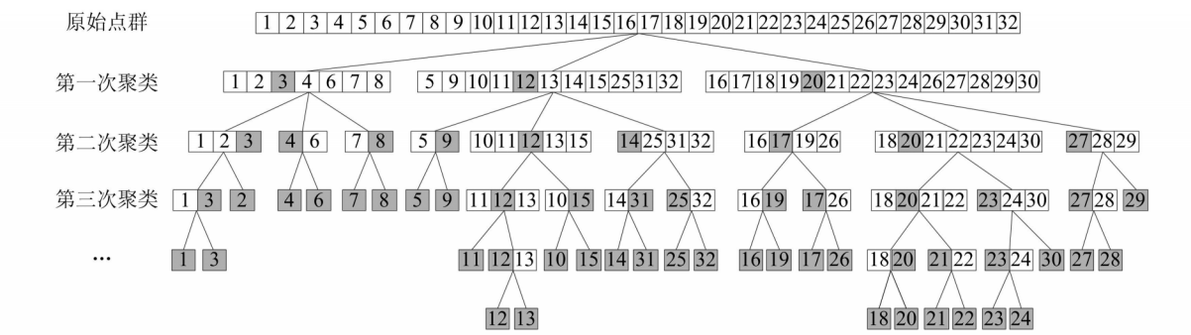


图5 点群的层次树结构

Fig.5 Hierarchical Trees of Point Group

按照上述步骤,即可获得满足数量和密度对比要求的选取结果。需要说明的是,在步骤3)中将选取点数分配到各簇时,由于会出现选取数量为非整数的情况,为了简化分配过程,可按照四舍五入取整的方法获得各类簇需要选取的数量,最终选取数量与开方根定律计算的结果可能会有少量偏差。当然也可以根据小数点后面的数字大小分配开方根计算的选取数量,得到严格满足开方根定律的选取结果。

以第1层级聚类为例,设簇数量为k,各簇点数为n_j($j=1,2\cdots k$),则各簇应选取的数量为:

$$N_j=\mathrm{int}(N_{\mathrm{select}}\cdot n_j/n) \tag{12}$$

式中,int(·)表示四舍五入取整。同理对于其他级别点簇,可得到相应的选取数量。

继续上述算例,假设需要选取18个点,为了便于描述,用二元对(L-N)表示簇的编号,其中$L=1,2,3\cdots$表示聚类的层级,$N=1,2,3\cdots$表示簇的序号。则第1次聚类结果中,簇(1-1)、(1-2)、(1-3)应选取点数为4、6、8;将3个簇中应选取的点数向下一级别簇分配,以簇(1-1)为例,簇(2-1)、(2-2)、(2-3)应选取点数为2、1、1;以此类推,簇(1-1)最终选取结果点编号为3、2、4、8。同理,原始点群的最终选取结果点编号为3、2、4、8、9、11、12、15、31、25、19、17、20、21、23、30、27、29,如图6所示。

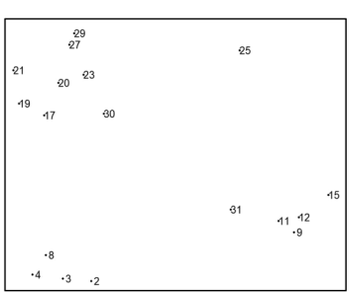


图6 图2(a)的选取结果

Fig.6 Result of Fig.2(a)

3 实验与分析

3.1 不同分布模式点群选取实验

实验由点群的自动聚类、层次树构建及点群的自动选取3部分组成,整个过程在Visual Studio 2010开发环境中基于ArcGIS Engine 10.2开发实现。采用Vmap1格式数据,比例尺为1∶25万,从可以公开下载的数据中选取美国3个地区不同分布模式的3类数据,分别是聚集分布的工业抽象点数据(对应图7中结果1~3)、均匀分布的高程点数据和随机分布的城市点数据,具体见表1。分析发现属性项中没有完整表征3类数据中各点重要性的等级信息,因而不能简单地依据属性信息进行选取。运用本文方法,分别进行1∶50万、1∶100万、1∶400万制图综合操作,选取点数见表2,原始点群及选取结果如图7所示(未按比例尺绘制)。从综合结果来看,不同数量的选取结果形成了从详细到概略的多层次表达,3种分布类型的数据都能按照指定的数量获得相应的选取结果,且均能较好地保持原始点群的空间分布形态。为了对选取结果进行定量评价,利用3类数据初次聚类中心对图幅进行Voronoi剖分,分别划分出14、27、16个综合区,按图8所示对综合区进行编号。统计各多边形综合区内各比例尺数据点数,以综合区编号为横轴、所在综合区点数为纵轴,绘制3类数据各综合区各次选取点数折线图,如图9所示。

表1 实验数据

Tab.1 Experimental Data

图幅名称	图层名称	分布模式
N091104	Industry ExtractPoints	聚集分布
N121201	SpotElevations	均匀分布
N111716	CityPoints	随机分布

表 2　实际选取的数量

Tab.2　Actual Selected Number

比例尺	数据种类	工业抽象点	高程点	城市点
1∶25万	原始点数	395	327	570
1∶50万	开方根定律	279	231	403
	实际选取	277	230	408
1∶100万	开方根定律	198	163	285
	实际选取	209	161	291
1∶400万	开方根定律	99	82	143
	实际选取	103	90	144

从图 9 可以看出,各综合区不同比例尺综合结果与原始点群在数量上基本保持一定的比例关系。对于各比例尺的综合结果,密集的地方依然密集,稀疏的地方依然稀疏,说明本文方法较好地保持了密度对比要求。从图 9 中还可看出,原始点群、1∶50 万、1∶100 万及 1∶400 万综合结果 4 条折线没有出现交叉,表明利用本文方法进行点群选取没有出现违反密度对比要求的情况。

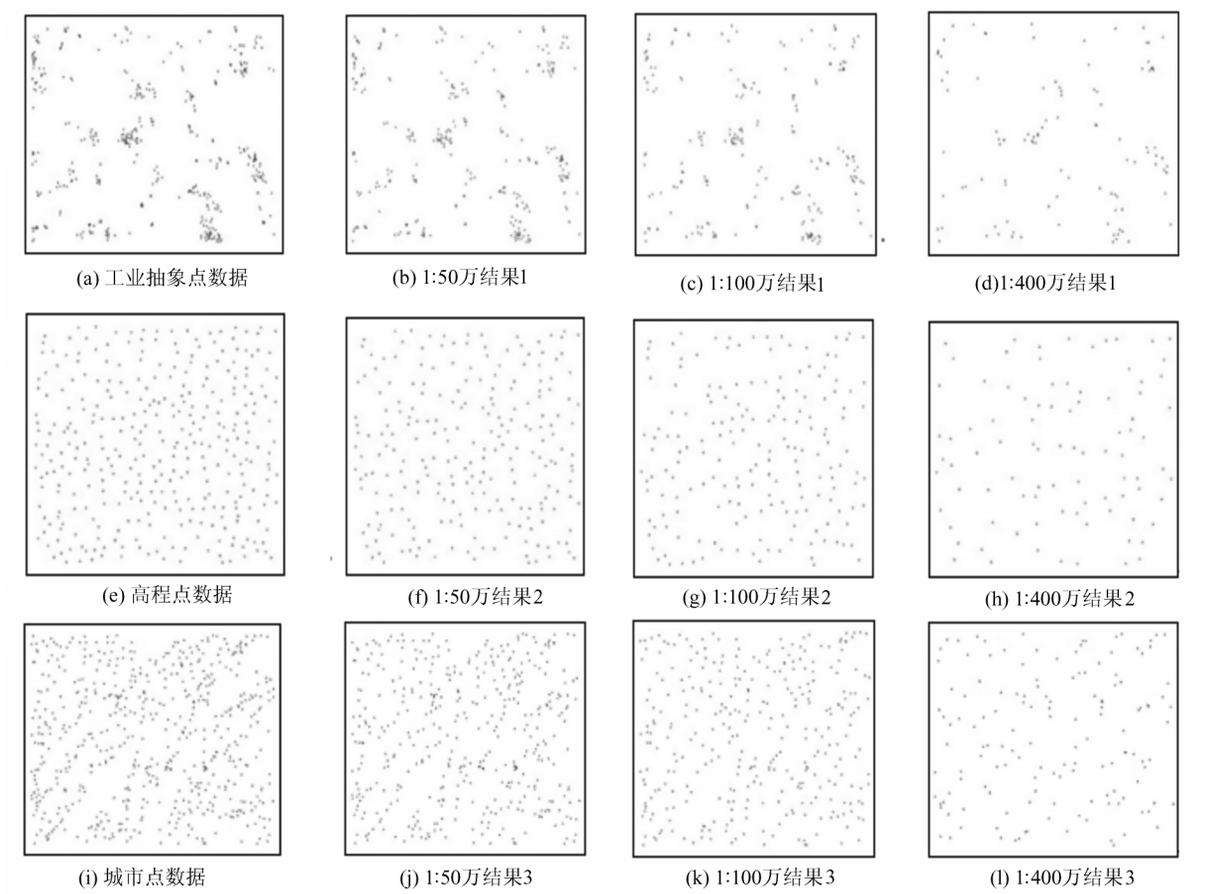


图 7　不同分布模式点群综合结果(未按比例尺绘制)

Fig.7　Point Group Generalization Results with Different Distribution Patterns

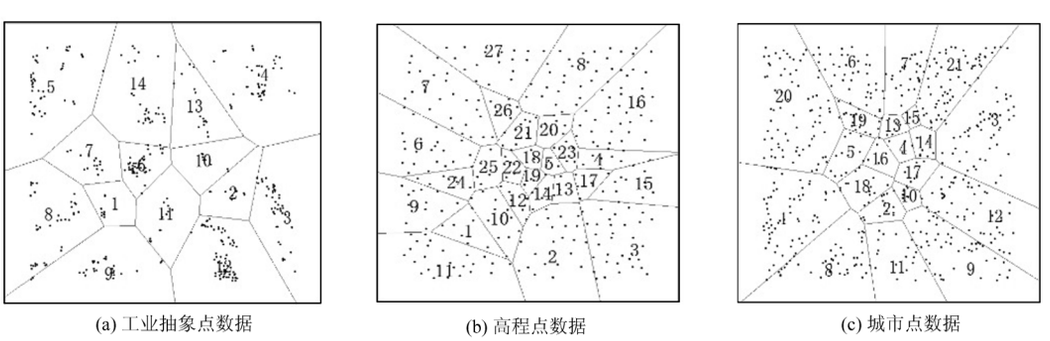
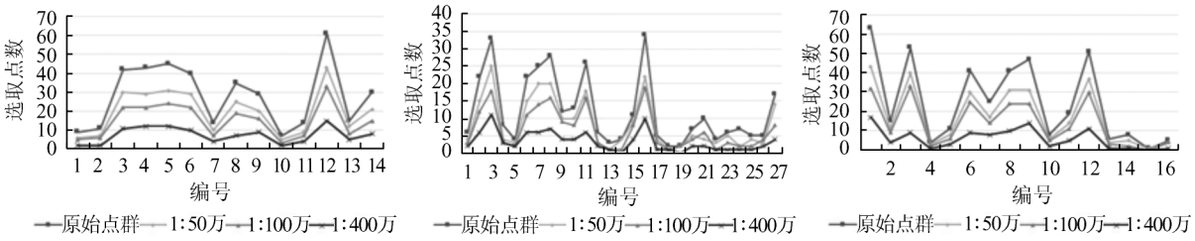


图 8　综合区划分示意图

Fig.8　Schematic Diagram of Generalization Area Division



(a) 工业抽象点数据 (b) 高程点数据 (c) 城市点数据

图9 各综合区选取点数对比

Fig.9 Comparison of the Numbers of Selected Points

3.2 与现有方法的比较分析

按照文献[19]中基于信息传递的点群综合评价方法,本文方法直接考虑了点群的统计信息(按开方根定律确定选取数量)和度量信息(保持密度对比),间接考虑了点群的拓扑信息(只选取聚类中心点),未直接考虑专题信息(语义信息),可见本文方法主要针对语义信息缺乏的情况。在语义不完备的情况下,即等级信息不完整或者部分要素没有语义信息时,可将所有要素一起聚类,按照本文方法构建多层次树,对于等级高的要素,无论其在层次树的哪个等级,直接标注选取,然后在各级别簇中选取剩余数量的点,实现选取。本文相比文献[4]的优势主要是可选取任意数量的点,且满足制图综合中的密度对比要求。

4 结 语

本文提出了一种顾及密度对比的多层次聚类点群选取方法,通过利用聚类中心表达相应层级的点簇,保证选取结果与原始点群空间分布形态上的整体一致;根据点群层次树结构逐层向下分配待选取点数,保证满足密度对比要求。实验验证了本文方法的可行性和普适性。本文方法的优点一是无需依靠点群的属性信息即可实现点群的多尺度表达;二是选取数量完全可控。本文的点群选取方法本质上只利用了点群间的距离关系,如何综合考虑点群间的方向关系和拓扑关系,并通过聚类的方法实现更科学合理的选取是下一步需要研究的内容。

参 考 文 献

[1] Yan Haowen, Wang Jiayao. A Generic Algorithm for Point Cluster Generalization Based on Voronoi Diagrams[J]. *Journal of Image and Graphics*, 2005, 10(5):633-636(闫浩文, 王家耀. 基于Voronoi图的点群目标普适综合算法[J]. 中国图象图形学报, 2005, 10(5):633-636)

[2] Guo Qingsheng, Zheng Chunyan, Hu Huake. Hierarchical Clustering Method of Group of Points Based on the Neighborhood Graph[J]. *Acta Geodaetica et Cartographica Sinica*, 2008, 37(2):256-261(郭庆胜, 郑春燕, 胡华科. 基于邻近图的点群层次聚类方法的研究[J]. 测绘学报, 2008, 37(2):256-261)

[3] Yan Haowen, Wang Bangsong. A MWVD-Based Algorithm for Point Cluster Generalization[J]. *Geomatics and Information Science of Wuhan University*, 2013, 38(9):1 088-1 091(闫浩文, 王邦松. 地图点群综合的加权Voronoi算法[J]. 武汉大学学报·信息科学版, 2013, 38(9):1 088-1 091)

[4] Li Jiatian, Kang Shun, Luo Fuli. Point Group Generalization Method Based on Hierarchical Voronoi Diagram[J]. *Acta Geodaetica et Cartographic Sinica*, 2014, 43(12):1 300-1 306(李佳田, 康顺, 罗富丽. 利用层次Voronoi图进行点群综合[J]. 测绘学报, 2014, 43(12):1 300-1 306)

[5] Li Wenjing, Li Shaoning, Long Yi, et al. Point Cluster Selection in GIS Using Gravity Model[J]. *Geomatics and Information Science of Wuhan University*, 2013, 38(8): 945-949(李雯静, 李少宁, 龙毅, 等. 利用重力模型进行GIS点群选取[J]. 武汉大学学报·信息科学版, 2013, 38(8): 945-949)

[6] Yan H, Weibel R, Yang B. A Multi-parameter Approach to Automated Building Grouping and Generalization[J]. *GeoInformatica*, 2008,12(1):73-89

[7] Steinhauer J H, Wiese T, Freksa C, et al. Recognition of Abstract Regions in Cartographic Maps[J]. *Lecture Notes in Computer Science*, 2001, 2 205: 306-321

[8] Jain A K. Data Clustering: 50 Years Beyond *k*-means[C]. European Conference on Machine Learning and Knowledge Discovery in Databases, Antwerp, Belgium, 2008

[9] Sander J. Density-Based Clustering[M]. US: Springer, 2011

[10] Achtert E, Goldhofer S, Kriegel H P, et al. Evaluation of Clusterings—Metrics and Visual Support[C].

IEEE 28th International Conference on Data Engineering (ICDE), Washington D C, USA, 2012

[11] Xu R, Wunsch D. Survey of Clustering Algorithms [J]. *IEEE Transactions on Neural Networks*, 2005, 16(3): 645-678

[12] Guha S, Rastogi R, Shim K. CURE: An Efficient Clustering Algorithm for Large Databases[J]. *ACM SIGMOD Record*, 1998, 27(1):73-84

[13] Kalita H K, Bhattacharya D K, Kar A. A New Algorithm for Ordering of Points to Identify Clustering Structure Based on Perimeter of Triangle: Optics (bopt)[C]. The 15th International Conference on Advanced Computing and Communications, Guwahati, India,2007

[14] 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, USA, 1998

[15] Rhouma M B H, Frigui H. Self-Organization of Pulse-Coupled Oscillators with Application to Clustering [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001,23(2):180-195

[16] Rodriguez A, Laio A. Clustering by Fast Search and Find of Density Peaks[J]. *Science*, 2014, 344 (6 191):1 492-1 496

[17] Xie J, Gao H, Xie W, et al. Robust Clustering by Detecting Density Peaks and Assigning Points Based on Fuzzy Weighted *k*-nearest Neighbors[J]. *Information Sciences*, 2016, 354:19-40

[18] Li D, Wang S, Gan W, et al. Data Field for Hierarchical Clustering [J]. *International Journal of Data Warehousing and Mining*, 2011,7(4):43-63

[19] Yan H, Weibel R. An Algorithm for Point Cluster Generalization Based on the Voronoi Diagram [J]. *Computers and Geosciences*, 2008, 34(8):939-954

A Point Group Selecting Method Using Multi-level Clustering Considering Density Comparison

CHENG Mianmian[1]　*SUN Qun*[1]　*LI Shaomei*[1]　*XU Li*[1]

1　Institute of Geospatial Information, Information Engineering University, Zhengzhou 450001, China

Abstract: In the absence of semantic information, the selection of point group is one of the difficulties in cartographic generalization. This paper proposes a new multi-level clustering point group generalization method which takes into account density contrast. Firstly, in view of the shortcoming of *k*-means clustering algorithm, this paper uses an improved density peak clustering method to realize automatic clustering of point group, mainly reflects on determining the optimal cut-off distance by the Gini coefficient and uses the relation of local density and relative distance to detect the clustering centers. Secondly, we propose a point group selection strategy which takes into account density contrast, the point group is divided into clusters of different grade by multi-level clustering. The clustering centers of different grades are determined, and the hierarchical tree structure of point group data is established. The number of points to be selected is calculated according to the square root law, then allocated from top to bottom according to the number of clusters at each level, and the selected objects are determined, and automatic selection of points and multi-scale expression of point group are realized. Point groups experimental results with different distribution patterns show that the method described in this paper can get reasonable selection results, which verifies the universality and effectiveness of the method.

Key words: spatial clustering; density peaks; Gini coefficient; point group selection; cartographic generalization

First author: CHENG Mianmian, PhD candidate, majors in multi-source spatial data fusion and cartographic generalization. E-mail:chmmian@163.com

Corresponding author: SUN Qun, PhD, professor. E-mail: sunqun@371.net

Foundation support: The National Natural Science Foundation of China, No. 41571399.