文章编号: 1671-8860(2010)12-1496-04 文献标志码: A

利用 k 阶空间邻近图的空间层次聚类方法

宋晓眉[1] 程昌秀[1] 周成虎[1] 陈荣国[1]

(1 中国科学院地理科学与资源研究所资源与环境信息系统国家重点实验室, 北京市大屯路甲11号, 100101)

摘 要: 分析了图斑 k 阶邻近图的特征及其层次关系, 并根据 k 阶空间邻近图对层次聚类方法进行改进, 不但使得聚类的时空信息的结合灵活有效, 而且降低了算法的时空复杂度。实验证明, 基于 k 阶空间邻近图的层次聚类具有较好的空间聚类效果。

关键词: k 阶空间邻近; k 阶空间邻近图; 空间层次聚类; 距离矩阵

中图法分类号: P208

在空间层次聚类中, 空间信息与属性信息的结合是一个令人头痛的问题[1-3]。在一定意义上, 空间层次聚类是在传统层次聚类的基础上添加了空间约束条件, 这种空间约束条件可反映到数据(或者类簇)相互关系的访问上, 因而实现对属性的结合。空间层次聚类相互关系的访问范围仅局限于当前类的周围几个相邻的类, 因此, 其算法在空间和时间上优于传统层次聚类。本文引进 k 阶空间邻近图的概念, 使空间邻近关系要求更加灵活, 同时采用一维数组的距离值存储结构以及相关的更改和移除一维数组更新操作, 在保证空间关系的实时调整的同时节省了存储空间, 也在一定程度上节省了时间。

1 k 阶空间邻近与 k 阶空间邻近图

1.1 图斑的 k 阶空间邻近

在文献[4, 5]的基础上, 本文给出了基于连续图斑邻接关系网的 k 阶空间邻近的定义。设 P_1、P_2 是邻接网中的两个图斑的几何中心点, 如果这两个点最少经过 k 个邻接关系网的边连通, 则称这两个图斑之间的邻近距离为 k, 记为 $d\mathrm{Neighbor}(P_1, P_2) = k$。称这两个图斑之间的关系为 k 阶空间邻近关系, 记为$< P_1, \mathrm{Neighbor}(k), P_2 >$。即

$$< P_1, \mathrm{Neighbor}(k), P_2 > <=> d\mathrm{Neighbor}(P_1, P_2) = k$$

k 阶空间邻近有性质: ①对称性, 即 $d\mathrm{Neighbor}(P_1, P_2) = k$ 与 $d\mathrm{Neighbor}(P_2, P_1) = k$ 等价。②拓扑度量性。在拓扑变化下, 邻近阶数不会发生变化。③唯一性。即使发生了拓扑变化, 每个图斑的空间邻近关系并不会发生变化。④递归性。某个图斑的 m 阶邻近图斑可以这样递归定义: $m-1$ 阶邻近图斑的1阶邻近图斑除去不大于 $m-1$ 阶图斑的所有图斑。

1.2 k 阶空间邻近图

连续分布的图斑集里, 任意图斑 P_1 和 P_2 满足 $d\mathrm{Neighbor}(P_1, P_2) \leqslant k$, 那么, 两者之间建立关联关系, 这些关系的集合称为图斑的 k 阶空间邻近图。k 阶邻近图斑具有对称性, 空间 k 阶邻近图的创建实际上是创建一个无向图, 因此需要遵守一定的法则, 否则采用笛卡尔遍历的方式将会产生一些重复数据, 在创建空间关系结束时还要重新遍历去除重复数据。这样不仅增加了时间复杂度, 对空间复杂度也提出了不必要的要求。利用对象的固有ID字段, 合理设计 k 阶邻近图建立的算法流程, 既可以节省临时存储空间又可以减少时间消耗。

假设有 n 个空间数据对象, OBJECTID 是从0到 $n-1$ 的。建立空间关系是按 OBJECTID 从小到大遍历, 对于某个图斑搜寻不大于 k 阶的邻近图斑, 如果搜寻图斑的 OBJECTID 大于当前图

收稿日期: 2010-11-18。

项目来源: 中国科学院知识创新工程重要方向资助项目(kzcx2-yw-304); 国家863计划资助项目(2007AA120401); 国家科技支撑计划资助项目(2007BAH16B03)。

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

斑则建立邻近关系。

2　基于 k 阶空间邻近图的空间层次聚类方法

2.1　距离矩阵弊端分析

传统层次聚类算法每次合并完一个类后, 就必须重新计算合并后新的类与旧的类的距离, 也就是距离矩阵。假定聚类的数据有 N 个, 那么按照传统层次聚类算法的思想, 第一次合并之前距离矩阵大小为 $N \times N$, 当合并完一个类之后, 距离矩阵通过收缩(两行和两列)和扩张(一行和一列)操作, 距离矩阵的大小变为$(N-1) \times (N-1)$。在第二次合并之前必须重新获取类与类之间的最小距离, 如果不采用特殊的存储结构或者索引结构[6], 最小距离的获取只能通过遍历, 遍历的大小为$(N-1) \times (N-1)$。依次递推, 直至所有类合并为一个类为止。传统层次聚类算法所需的距离矩阵存储空间是 N^2 阶, 而每次类的合并又要重新获取最小距离值, 相应计算量为 N^2 阶。因此, 传统的层次聚类的时间复杂度是 $o(N^3)$。

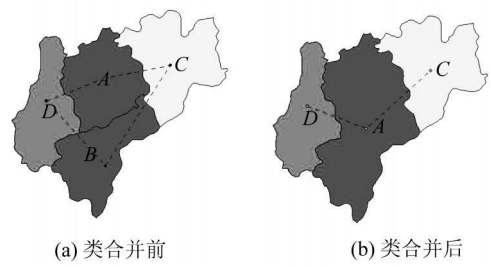
矩阵的收缩操作会将两个类的所有信息删除, 以便实现扩张操作对距离矩阵的更新。但是对于空间层次聚类, 所有信息的删除意味着原有的空间信息亦被删除, 扩张操作需要对新生成的类的空间信息进行重新确定。图斑的空间邻近关系的确定是一个复杂的过程, 尤其是在类的合并过程中。类的合并在空间意义上是指两个图斑的图形合并, 还需要重新确定新图斑的空间关系。

因此, 对于空间层次聚类需要明确: 空间关系的判断不应多次被调用, 新类与旧类空间关系的判断需要依据初始空间关系来判断, 否则, 在聚类过程中会增大时空复杂度和不确定性; 空间数据的空间信息反映在聚类中该不该进行距离值的计算上, 这个考虑应该贯彻在整个聚类的过程中, 传统的距离矩阵不再适合作为存储结构。本文的空间层次聚类时间和空间复杂度目标是在不考虑初始空间关系判断的基础上不大于传统空间层次聚类。

2.2　存储结构以及相关操作的设计

本文采用简单的一维数组作为存储结构, 设计距离结构体保存空间信息与类的属性信息, 并采用更新与移除操作, 以满足类的合并过程中属性信息与空间信息的同步更新。不能对关系直接进行简单的删除操作, 只能采取更新操作。这样可以保留空间信息不被破坏, 同时, 又可以更改新类的属性值, 相当于更新新类与其他类的距离。在找到最小距离值后, 进行类的合并是通过对类的距离结构体进行更新以及重复距离结构体的移除操作来实现的。涉及到的结构体中包含两个类的 ID 号, 并且只包含两个中的一个 ID 号。对于包含 ID 大的距离结构体, 需要进行 ID 号更改为小 ID 号的操作, 并且对 Value 和 Num 的属性值更新; 对于包含 ID 小的距离结构体, 只需要对 Value 和 Num 的属性值更新。

在进行 ID 号的修改的时候一般情况下会产生重复邻近关系。如图 1 所示, A 和 B 将要合并为一类, 而 A 和 B 同时分别和 C 和 D 存在空间邻近关系, 显然会存在 4 个距离结构体 Struct_Dist(AD)、Struct_Dist(AC)、Struct_Dist(BD)和 Struct_Dist(BC)。假设 A 的 ID 号小于 B 的 ID 号, 那么 A 只需要进行属性值更新, B 需要改 ID 号为 A 的 ID 号并进行属性值更新。这样, 原来的 AB 就变成完全一样的 A, 实现了类的合并。但是, 这样也产生了两条 Struct_Dist(AD) 和 Struct_Dist(AC)的关联关系结构, 这需要进行移除操作。移除操作是在找到关联距离结构体之后, 对结构体进行更新之前进行。找到一个关联结构体, 先看看如果修改结构体, 是否和前面的修改过的结构体出现重复, 如果会重复就将此结构体移除, 否则进行更新操作。



(a) 类合并前　　(b) 类合并后

图 1　处理重复结构体

Fig. 1　Dealing with Repeated Structures

聚类结果树采用多叉树结构存储。每次合并类, 将较小类 ID 号指向较大类 ID 号, 同时, 修改距离中相应较大的 ID 为较小的 ID 号。在聚类结束时, 可以通过一维距离数组中剩下的结构体递归寻找该类中的类成员。

2.3　空间层次聚类流程

本文方法的思想是, 在已经建立空间关系(空间邻近关系)文件的基础上, 采用一维距离数组存储结构和相应的距离结构体的更新移除操作, 实现类的合并, 通过递归最终实现空间层次聚类。具体算法流程为: ①读取 k 阶邻近图。②类归并过程。③递归合并类。每进行一次类的合并就意味着类

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

数减少 1, 因此假设图斑个数为 n, 凝聚式聚类的数目为 m 时, 需要进行 $n-m$ 次类的归并即可。

3 实验与分析

本文对山东省内陆县域的 55 a 降水累计值的平均值进行空间层次聚类, 要求聚类结果在属性值相似的前提下在空间上保持邻近性。本文实验要求空间邻近关系满足 1 阶邻近和 2 阶邻近, 最后与传统的层次聚类结果进行比较分析。

3.1 数据来源

数据采用山东省内陆县域(108 个, 剔除了长岛县, 因为它由离散图斑组成, 与其他的县域没有邻接关系), 东营市辖区包含一块飞地。图形数据是使用的图斑, 属性数据采用的是降水累计值 55 a 的平均值的取整。如图 2 所示, 图中标注的就是属性值。

3.2 建立空间邻近图

对示例数据建立空间邻近关系, 使用 xml 文件存储。可视化结果如图 3 所示, 黑色线段连接两个图斑表示该两图斑存在邻近关系, 二阶邻近图的连接线明显比一阶的稠密(图中最上方一个图斑表示的是东营市辖区, 图斑的中心向下偏移, 因为它下方存在一块飞地。在视觉上飞地造成显示的关联线段相交, 但这并不影响后面聚类的正确执行。)

3.3 聚类结果

建立邻近图并存储在文件中可以方便以后其他属性数据的空间层次聚类。读取 k 阶空间邻近图文件, 设置聚类的类数为 4, 开始空间层次聚类建立聚类结果树。递归读取结果树, 获取聚类结果, 如图 4 所示。

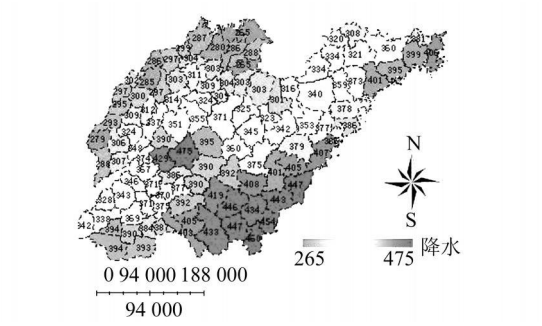


图 2 山东主要县域平均总降雨量分布图

Fig. 2 Distribution of Average Total Rainfall of Main Counties in Shandong

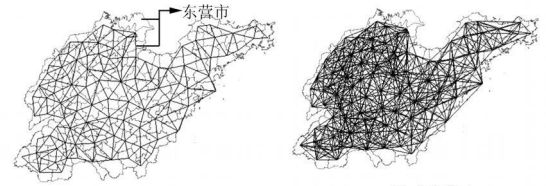
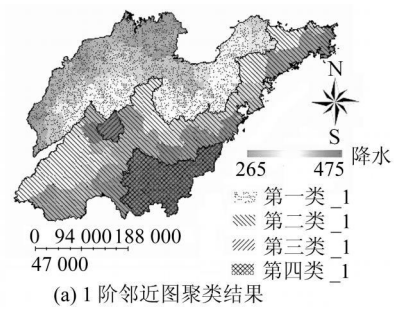


图 3 空间邻近图

Fig. 3 k-order Spatial Neighboring Map

3.4 结果分析

本文采用的空间数据通过实验可以得知图斑的最大阶是 16, 就是说不小于 16 阶的邻近图是完全一样的, 每个空间数据与其他数据都存在空间邻近关系, 这时的空间层次聚类就退化为传统的层次聚类。进行 17 阶邻近图的空间层次聚类结果如图 5 所示。



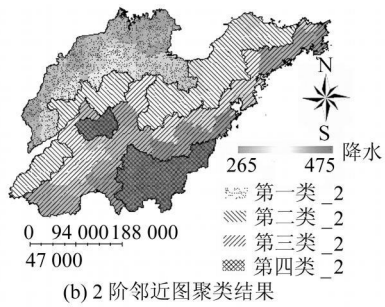


图 4 基于邻近图聚类结果

Fig. 4 Clustering Results Based on Neighboring Map

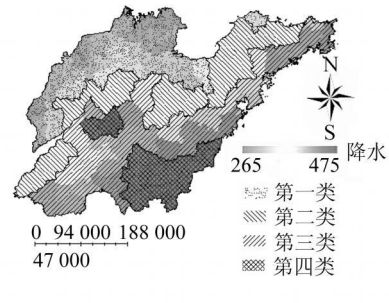


图 5 传统层次聚类结果

Fig. 5 Results of Traditional Hierarchical Clustering

图 4(a)中 1 阶邻近图的层次聚类结果具有最强的空间邻近性, 保证了所有的类成员之间是相互连接的, 没有分离的成员存在。第三类的属性值与第四类的属性值比较接近, 但是由于空间的原因而不能划分为一类; 第三类的属性值与第一类和第二类的属性值相差很大, 尽管相邻近但是仍然不能划分为一类。对比图 5 中的传统层次聚类结果可以看出, 1 阶邻近图的层次聚类结果使分布离散的数据更加集中, 类内的属性相似性有所降低, 但是空间分布上具有最大的相似性。

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

图 4(b) 中 2 阶邻近图的层次聚类结果放松了空间邻近性要求, 致使 1 阶邻近图的层次聚类结果第四类以及第二类的一个图斑划分到第四类中, 这和传统层次聚类的结果是一样的。另外, 第三类的结果也与传统聚类的聚类结果一致, 只有第一类与第二类的结果由于空间要求的限制存在少许的不同。可见, 2 阶邻近图的层次聚类可以很好地反映属性的聚集性, 并且在空间上保持很好的集聚性。

本文采用的 k 阶邻近图约束的空间层次聚类, k 的取值不同, 可以使聚类结果不同。k 值越小(最小为 1), 空间要求越高, 属性的相似性要求就会最低; k 值越大, 空间要求越低, 属性的相似性要求就会越高。当 k 值大到一定程度之后, 空间层次聚类就会退化为传统层次聚类, 这时采用一维距离数组存储结构并不会显示优势, 但是聚类过程中时间复杂度并不会比传统层次聚类差, 因为二者的效率瓶颈都是最小距离值的查找。

4　结　语

空间层次聚类思想虽然在很大程度上雷同于传统层次聚类, 但是又有自己的独特之处: 聚类过程中对数据空间关系的约束减少了数据之间的相互计算。比如, 在层次树的计算时, 数据的距离计算仅限于与它相邻的少数几个数据, 从而降低了空间复杂度。此外, 本文所述方法的时间复杂度为 $o(N^2)$, 传统层次聚类算法的时间复杂度为 $o(N^3)$, 因此, 此方法的时间复杂度较低。

参　考　文　献

[1] 张燕文. 基于空间聚类的区域经济差异分析方法[J]. 经济地理, 2006, 26(4): 557-560

[2] 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

[3] 王海军, 张德礼. 基于空间聚类的城镇土地定级方法研究[J]. 武汉大学学报·信息科学版, 2006, 31(7): 628-631

[4] 闫超德, 白建军, 赵仁亮. 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, 2009, 34(1): 48-51

[5] 杜晓初, 郭庆胜. 基于 Delaunay 三角网的空间邻近关系推理[J]. 测绘科学, 2004, 29(6): 65-67

[6] 张振亚, 程红梅, 王进, 等. 面向凝聚式层次聚类算法实现的矩阵存储数据结构研究[J]. 计算机科学, 2006, 33(1): 14-17

第一作者简介: 宋晓眉, 博士生, 主要从事空间聚类、空间数据库查询优化研究。

E-mail: songxm@lreis.ac.cn

Spatial Hierarchical Clustering Method Based on *k*-order Spatial Neighboring Map

SONG Xiaomei[1]　*CHENG Changxiu*[1]　*ZHOU Chenghu*[1]　*CHEN Rongguo*[1]

(1　LREIS, Institute of Geographical Sciences and Natural Resources Research, CAS, A11 Datun Road, Beijing 100101, China)

Abstract: In the spatial hierarchical clustering, the effective integration of attribute information and spatial information is a very important issue. We analyze the characteristics and hierarchy of the k-order spatial neighboring map, and improve the hierarchical clustering based on k-order spatial neighboring map. Therefore, the algorithm not only makes the combination of spatial and attributes information flexible and effective, but also makes the time and space complexity of this method lower than the traditional algorithm. Finally, the experimental results prove that spatial hierarchical clustering based on k-order spatial neighboring relations has good spatial clustering effect.

Key words: *k*-order spatial neighboring relations; *k*-order spatial neighboring map; spatial hierarchical clustering; distance matrix

About the first author: SONG Xiaomei, Ph. D candidate, majors in spatial clustering and query optimization of spatial database.

E-mail: songxm@lreis.ac.cn

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net