第32卷 第21期 计 算 机 工 程 2006年11月
Vol.32 No.21 Computer Engineering November 2006
·博士论文· 文章编号：1000—3428(2006)21—0050—02 文献标识码：A 中图分类号：TP312

一种基于空间邻接关系的k-means聚类改进算法

王海起[1,2,3]，王劲峰[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 100101；
2. 中国石油大学(华东)地球资源与信息学院，东营 257061；3. 中国科学院研究生院，北京 100039)

摘 要：空间对象不仅具有非空间的属性特征，而且具有与空间位置、拓扑结构相关的空间特征。利用传统的聚类方法对空间对象进行聚类时，由于没有考虑空间关系，同一类的对象可能出现在空间不相邻的位置。基于空间邻接关系的k-means改进算法将相邻对象的空间邻接关系作为约束条件加以考虑，使聚类结果既反映了属性特征的相似程度，又反映了对象的空间相邻状态，从而可以揭示不同类别对象的空间分布格局，因此其比传统的k-means方法更适合于空间对象的聚类分析。

关键词：空间对象；空间邻接关系；邻接矩阵；k-means聚类算法

A k-means Adapted Algorithm Based on Spatial Contiguity Relations

WANG Haiqi[1,2,3], WANG Jinfeng[1]
(1. National Key Lab of Resources and Environment Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101; 2. College of Geo-resources and Information, University of Petroleum (East China), Dongying 257061; 3. Graduate School of Chinese Academy of Sciences, Beijing 100039)

【Abstract】 Spatial object has not only non-spatial attribute properties but also spatial properties related with space coordinates and topological structures. When using the traditional clustering methods to classify spatial objects, the objects of the same class may appear in non-adjacent spatial positions because spatial relationships are not been considered. The k-means adapts algorithm based on spatial contiguity relations regards spatial contiguities of the neighboring objects as a restrained condition. So the clustering result not only reflects the similarities of attributes but also reflects spatial adjacent relations, and furthermore reviews spatial distribution patterns of different classes. Therefore, this adapted algorithm is more suitable for the clustering analysis of spatial objects than the traditional k-means method.

【Key words】 Spatial object; Contiguity relation; Contiguity matrix; k-means clustering algorithm

1 概述

聚类方法是数据挖掘领域中应用极其广泛的技术，它基于“物以类聚”的思想，将研究对象按照特征分组为多个类，每个类对象之间具有较高的相似度，而不同类对象之间的差别较大。相似度可以基于距离、密度或模型等来度量。Han等把常用的聚类方法分为如下几类：分割法(Partitioning)，分层法(Hierarchical)，基于密度的方法(Density-based)，基于网格的方法(Grid-based)和基于模型的方法(Model-based)[1]。k-means算法属于基于分割的方法，由MacQueen提出，是目前应用最为广泛的一种聚类方法。

在GIS空间数据分析(Spatial Data Analysis)中，空间对象主要指属性数据关联于具有规则或不规则边界的多边形区域的对象，也称区域对象或lattice对象[2]。不同对象间的空间关系按照拓扑结构可表达为空间邻接矩阵，即当2个空间对象具有共同边界或共享一个顶点时，则认为这2个对象在空间上是相邻的，相应的邻接矩阵元素取值为1；否则，取值为0。

利用聚类方法对空间对象进行分类时，由于空间关系的存在，要求同一类中的对象在空间上处于相邻的位置，在地图上表现为彼此相连的状态。而传统的聚类方法仅利用空间对象的属性数据，并没有考虑对象的空间邻接关系。有研究在对空间对象聚类时，将对象的空间坐标作为额外的属性变量加以考虑，然而这种方法得到的同一类的对象仍然可能出现在空间不相邻的位置[3]；也有研究提出了一些新的空间对象聚类方法[4]。

本文利用k-means聚类方法对空间区域对象进行分类，在聚类过程中将空间邻接关系作为约束条件加以考虑。在对每个空间对象进行类别归属判断时，不仅要考虑对象与某类别中心的距离，而且要考虑对象与该类别中已有空间对象的邻接关系；只有当该类别与进行归属判断的空间对象之间存在邻接关系且距离最短时，对象才可以归属于该类。这样，对于最终的分类结果，既保证了同一类内对象属性值差别较小、不同类之间属性值差别较大，又保证了同一类的对象在空间上处于相邻的位置。

2 基于空间邻接关系的k-means聚类算法

2.1 相关定义

在给出详细的基于空间邻接关系的k-means算法流程之前，首先对研究的GIS空间对象作如下定义：

(1)设研究区域S有N个空间对象$S=\{s_1, s_2,\ldots, s_N\}$及邻接关系(neighbor relation)$R\subseteq S\times S$。空间对象$s_i$和$s_j$具有邻接关系当且仅当$(s_i, s_j)\in R$，$i\neq j$。用空间邻接矩阵W表达邻接关系R，

基金项目：国家自然科学基金资助项目(40471111)；国家“863”计划基金资助项目(2002AA135230-1)；国家“973”计划基金资助项目(2001CB5103)

作者简介：王海起(1972－)，男，博士生，主研方向：GIS与空间信息分析；王劲峰，博士、研究员

收稿日期：2005-11-27 **E-mail：**wanghq@lreis.ac.cn

$W(i, j)=W_{ij}=1$ 当且仅当$(s_i, s_j)\in R$，否则$W(i, j)=W_{ij}=0$。

(2)对每个空间对象s_i，设对象的d维属性向量为$x_i\equiv x(s_i)=[x_{i1}, x_{i2},\ldots, x_{id}]$。

其次，对于 k-means 聚类算法作如下定义：

(1)定义$\{z_1, z_2,\ldots, z_K\}$为K个聚类中心，每个聚类中心$z_j=[z_{j1}, z_{j2},\ldots, z_{jd}]$，j=1, 2, …, K。

(2)对每个聚类中心z_j定义一个集合Z_j，用于存放该类别包含的空间对象，初始化时集合Z_j为空。

(3)定义 N×K 的二维距离矩阵 Dist，用于存放每个空间单元与每个聚类中心的距离。同时定义矩阵 Dist 的 N×K 辅助逻辑矩阵 DistMark，用于标识在距离矩阵 Dist 中搜索单元到聚类中心的最短距离时该距离是否参与搜索过程，若矩阵 DistMark 中某元素值为 True，则矩阵 Dist 中对应距离参与搜索，否则不参与搜索。

2.2 算法流程

基于空间邻接关系的 k-means 聚类详细算法步骤如下：

(1)利用GIS软件(ArcGIS、GeoDa等)构建空间邻接矩阵W[5]，并指定k-means聚类的类别数K；

(2)从N个空间对象中随机挑选K个对象作为初始聚类的各类别中心$\{z_1(0), z_2(0),\ldots, z_K(0)\}$；

(3)各类别的集合Z_j初始化为空，空间单元集合S初始化为包括所有空间对象，$S=\{s_1, s_2,\ldots, s_N\}$。

对于第m次迭代，分别计算每个空间对象s_i到每个聚类中心$z_j(m)$的距离$Dist(s_i, z_j)$，并将矩阵DistMark中各元素值赋为True，距离计算公式如下：

$$Dist(s_i, z_j^{(m)}) \equiv Dist(i, j)$$
$$= \left\| x_i - z_j^{(m)} \right\|^2 = \sum_{t=1}^{d} (x_{it} - z_{jt}^{(m)})^2 \quad i=1, 2,\ldots, N;\ j=1, 2,\ldots, K$$

(4)利用距离矩阵Dist和辅助矩阵DistMark搜索空间对象集合S中对象到聚类中心的最短距离minDist，参与搜索的距离对应的DistMark元素值必须为True，设搜索得到的最短距离minDist对应的空间对象为s_{min_i}，对应的聚类中心为z_{min_j}；

(5)若聚类中心z_{min_j}对应的集合Z_{min_j}为空，或对象s_{min_i}与集合Z_{min_j}中的空间对象具有邻接关系，即$W(s_{min_i}, Z_{min_j})=1$，则$S=S-\{s_{min_i}\}$，$Z_{min\_j}=Z_{min\_j}\cup\{s_{min\_i}\}$，并将矩阵DistMark中各元素值赋为True，继续步骤(6)。

若空间单元s_{min_i}与集合Z_{min_j}中的单元不具有邻接关系，令$DistMark(s_{min_i}, z_{min_j})$=False表示下次搜索不包括该距离，返回步骤(4)；

(6)若集合 S 不为空，返回(4)；为空，表明所有的空间对象已分别归属于 K 个不同的类别中，继续下一步；

(7)更新各聚类中心值$\{z_1(m+1), z_2(m+1),\ldots, z_K(m+1)\}$，计算公式如下：

$$z_j^{(m+1)} = (z_{j1}^{(m+1)}, z_{j2}^{(m+1)}, \ldots, z_{jd}^{(m+1)})$$
$$= \frac{1}{C(Z_j)} \sum_{s_i\in Z_j} x_i = \frac{1}{C(Z_j)} \sum_{s_i\in Z_j} (x_{i1}, x_{i2}, \ldots, x_{id}) \quad j=1, 2,\ldots, K$$

式中，$C(Z_j)$是各类别集合Z_j中包含的空间对象个数。

(8)若所有的聚类中心均保持稳定，即对j=1, 2, …, K，有$z_j(m)=z_j(m+1)$，则k-means聚类过程结束，各类别集合Z_j中包含了归属于各类的空间对象；否则，令m=m+1，返回(3)，继续下一次迭代。

3 应用实例

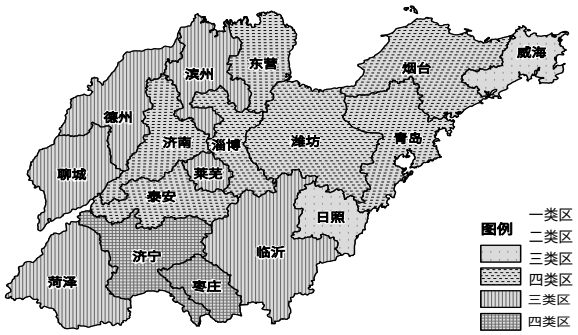
以山东省 17 个地区生态环境数据为例进行聚类分析。由于相邻地区的生态环境指标存在空间相关性，因此，在聚类过程中考虑空间邻接关系可以把握生态环境指标的空间分布形态和差异规律。表 1 为山东 17 个地区 2000 年生态环境指标数据[6]，图 1 和图 2 分别为传统k-means聚类结果和基于空间邻接关系的聚类结果。

表 1 山东 17 地区生态环境指标数据(2000 年)

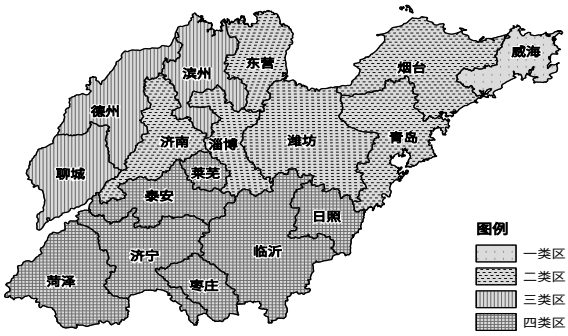
编码	地区	单位 GDP 废水排放量(t /万元)	废水达标率(%)	单位 GDP 废气排放量(m^3/ 元)	化肥施用密度(t / km^2)	农药施用密度(t/ km^2)	治理投资占 GDP 比重(%)
370100	济南	7.26	90.00	1.09	26.23	0.47	2.91
370200	青岛	7.92	96.53	0.76	29.30	0.80	1.73
370300	淄博	16.10	99.36	3.19	18.96	0.66	4.93
370400	枣庄	31.31	96.25	4.88	34.37	0.56	4.11
370500	东营	11.34	85.78	0.56	12.57	0.46	2.50
370600	烟台	8.22	97.52	1.18	22.83	1.35	1.89
370700	潍坊	13.78	90.26	1.23	32.55	1.14	2.14
370800	济宁	22.45	95.40	1.89	41.64	1.10	4.76
370900	泰安	18.26	83.77	1.67	22.73	0.82	4.07
371000	威海	4.35	9.59	0.65	16.25	1.78	0.42
371100	日照	13.20	8.73	1.58	25.70	1.32	0.44
371200	莱芜	15.75	3.56	5.50	17.42	0.58	1.77
371300	临沂	12.42	5.83	1.08	21.11	0.74	4.42
371400	德州	19.31	4.63	1.86	30.77	0.68	1.83
371500	聊城	22.29	7.49	0.84	35.71	0.97	6.[illegible]一类区
371600	滨州	11.58	8.42	0.46	26.02	0.63	4.[illegible]二类区
371700	菏泽	16.69	2.94	0.69	31.41	0.87	3.[illegible]三类区
							四类区

图 1 传统 k-means 聚类结果

图 2 基于空间邻接关系的 k-means 聚类结果



可以看出，仅考虑空间对象生态环境指标属性数据的传统 k-means 聚类结果存在同一类的对象在空间上处于不相邻位置的现象；而基于空间相邻关系的 k-means 聚类结果不仅刻画了生态环境指标数据的相似程度，而且刻画了各指标类别的空间分布格局，因此其聚类结果更为合理。



4 结论及讨论

GIS 空间对象既具有非空间的属性特征，又具有与位置相关的空间特征。基于空间邻接关系的 k-means 算法在对属性特征的聚类过程中考虑了不同对象的空间相邻性，从而使聚类方案可反映不同类别对象的空间分布状态。因此，这种

（下转第 75 页）