第 37 卷 第 16 期
Vol.37 No.16

计 算 机 工 程
Computer Engineering

2011 年 8 月
August 2011

·人工智能及识别技术· 文章编号：1000—3428(2011)16—0176—03 文献标识码：A 中图分类号：TP311.13

基于信息熵的空间对象群聚类算法

刘建兴[1]，鲍培明[1,2]

(1. 南京师范大学计算机科学与技术学院，南京 210097；2. 江苏省信息安全保密技术工程研究中心，南京 210097)

摘 要：针对利用空间关系建立空间对象群聚类的问题，提出一种基于信息熵的空间对象群聚类算法 ESOGC。该算法考虑空间数据的复杂性和数据之间的联系，根据邻域范围内信息熵的变化情况，捡起或放下当前空间对象群，从而实现对空间对象群的聚类。实验结果表明，该算法能解决空间对象群中对象类型、对象属性值和对象数量不一致性的问题。

关键词：空间对象群；空间关系；聚类；信息熵；蚁群算法

Clustering Algorithm for Spatial Object Group Based on Information Entropy

LIU Jian-xing[1], BAO Pei-ming[1,2]

(1. College of Computer Science and Technology, Nanjing Normal University, Nanjing 210097, China;
2. Jiangsu Research Center of Information Security and Confidential Technology Engineering, Nanjing 210097, China)

【Abstract】For the clustering of spatial object group constructed based on spatial relationship, this paper presents a clustering algorithm for spatial object group based on information entropy, named ESOGC. ESOGC is different from the other clustering algorithms, and it takes variety data types and the number of objects into full account in spatial object group. Through the change of information entropy within a same region, ants determine whether to pick up or drop the current spatial object group to realize the clustering of spatial object group. Experimental results show it can solve the problems of different data types, attribute value, and number.

【Key words】spatial object group; spatial relationship; clustering; information entropy; ant colony algorithm

DOI: 10.3969/j.issn.1000-3428.2011.16.060

1 概述

空间数据不同于一般的数据，它具有空间属性和非空间属性；空间数据之间有拓扑关系、方位关系和距离关系等；空间数据类型有点对象、线对象和面对象等。目前空间聚类[1-3]问题的解决方案尚局限在对点对象或单一对象的聚类，也没有考虑数据之间的空间关系。空间聚类的研究问题应扩展到点、线和面对象的聚类，以及空间对象群的聚类等。

空间数据集包含类型多样的空间对象，基于空间关系将空间数据集划分成若干个对象组，每一个对象组称为一个空间对象群[4]。2 个不同的空间对象群包含的对象类型、对象属性值和对象数量都可能不一样，例如，一个空间对象群包含的主要对象是居民区和商业网点，另一个空间对象群包含的主要对象是公园和河流。一个空间对象群中还可能同时包含有点对象、线对象和面对象。空间对象群的这些特性，使得空间对象群的聚类分析不能直接应用现有的空间聚类算法。

本文提出了基于信息熵的空间对象群聚类(ESOGC)算法，把空间对象群包含的对象看作空间对象群的属性，采用蚁群算法，根据空间对象群主题对象邻域范围内信息熵的变化实现空间对象群的聚类。

2 问题定义及概念

定义 1(空间对象群) D 是空间数据集，将 D 中对象分成 m 组 $\boldsymbol{S}_1,\boldsymbol{S}_2,\mathrm{L},\boldsymbol{S}_m$，$\boldsymbol{S}_i\subseteq D$，$i\in[1,m]$。称空间对象集合 $\boldsymbol{S}_i$ 为一个空间对象群。

空间对象群的形成是根据对象之间的方向、距离或拓扑关系等分析计算产生的，与应用问题相关。如基于面包含的拓扑聚类分析中，每一个面包含的对象集可以看作是一个空间对象群。空间对象的类型差异、属性值差异以及空间对象数目的差异，带来了空间对象群的复杂性。

定义 2(层) D 是空间数据集，将 D 中对象按类型分成 n 个子集 A_1,A_2,L,A_n，$D=\bigcup_{h=1}^{n}A_h$，$A_h\subseteq D$，$h\in[1,n]$，同一类型的对象具有相同的数据结构。每个空间对象子集 A_h 定义为一个空间对象层，简称层。

空间对象群和层是对空间数据集的不同划分，空间对象群是根据应用问题需求将空间数据集分组，层是按空间对象的数据结构将空间数据集分类。

将空间对象群 $\boldsymbol{S}_i$ 包含的对象分层表示后，2 个不同的空间对象群 $\boldsymbol{S}_i$ 和 $\boldsymbol{S}_j$ $(i\neq j)$ 的维数一致了。A_h 层上所有对象按属性值进行分类，假设分成 T 类 $A_h^1,A_h^2,\mathrm{L},A_h^T$，$A_h=\bigcup_{t=1}^{T}A_h^t$，$A_h^{t1}\cap A_h^{t2}=\varnothing$，$t1\neq t2$，$t1,t2\in[1,T]$。例如，若 A_h 层表示房屋，则 A_h 层对象可分类为高价房、一般价房和低价房等。每一层分类的数量与该层对象的属性值分布相关。

定义 3 空间对象群 $\boldsymbol{S}_i$ 包含的对象可能分属于不同的层上，用 B_{ih} 表示空间对象群 $\boldsymbol{S}_i$ 包含的 A_h 层的对象集，$B_{ih}=\boldsymbol{S}_i\cap A_h$，$i\in[1,m]$，$h\in[1,n]$，$B_{ih}$ 可以为空。$\boldsymbol{S}_i$ 可以用所包含每一层对象集的向量表示：$\boldsymbol{S}_i=(B_{i1},B_{i2},\mathrm{L},B_{in})$。

定义 4(空间对象群的主题对象) D 是空间数据集，根据

基金项目：国家自然科学基金资助项目(40871176)

作者简介：刘建兴(1986－)，男，硕士研究生，主研方向：空间聚类技术；鲍培明，副教授

收稿日期：2011-02-18 **E-mail：**jxliu_nnu@163.com

应用问题确定空间对象群 $\boldsymbol{S}_i$ 的主题对象，标为 O_i，$\boldsymbol{S}_i \subseteq D$，$O_i \in \boldsymbol{S}_i$，$i \in [1,m]$。

空间对象群主题对象的确定与应用问题相关，例如基于河流周边环境构建的空间对象群中，河流可以看作是空间对象群的主题对象。

定义 5(空间对象群主题对象的距离) 若空间对象群 $\boldsymbol{S}_i$ 和 $\boldsymbol{S}_j$ 的主题对象分别为 O_i、O_j，$i,j \in [1,m]$，则 $\boldsymbol{S}_i$ 和 $\boldsymbol{S}_j$ 的主题对象的距离定义为 $d(O_i,O_j)$，其中，$d(O_i,O_j)$ 为 O_i 和 O_j 的欧式距离。

定义 6(空间对象群主题对象的邻域) 若空间对象群 $\boldsymbol{S}_i$ 和 $\boldsymbol{S}_j$ 的主题对象分别为 O_i、O_j，$i,j \in [1,m]$，且 $d(O_i,O_j) < r$，则 $O_j \in N(O_i)$，$N(O_i)$ 表示 $\boldsymbol{S}_i$ 的主题对象 O_i 的邻域，r 为距离阈值。

定义 7(空间对象群主题对象的邻居) 若 O_j 为空间对象群 $\boldsymbol{S}_j$ 的主题对象且 $O_j \in N(O_i)$，则 $\boldsymbol{S}_i$ 与 $\boldsymbol{S}_j$ 为基于主题对象的邻居关系，其中，$i,j \in [1,m]$。

定义 8(信息熵) 假设空间对象属性相互独立，则 area 范围内所有空间对象群的信息熵按式(1)计算，其中，h 是层编号，t 是对象在 A_h 层上的类编号，$h \in [1,n]$，$t \in [1,T]$。p_{ht} 是 A_h 层上第 t 类对象的可能性函数，按式(2)计算，$|A_{ht}|_{\text{area}}$ 是 area 范围内 A_h 层上第 t 类对象的个数，$|A_h|_{\text{area}}$ 为 area 范围内 A_h 层上对象的个数。

$$E(\text{area}) = -\sum_{h=1}^{n}\sum_{t=1}^{T} p_{ht} \times \lg p_{ht} \tag{1}$$

$$p_{ht} = \frac{|A_{ht}|_{\text{area}}}{|A_h|_{\text{area}}} \tag{2}$$

对一个子空间来说，空间对象群相似时的信息熵小于空间对象群不相似时的信息熵。根据这个原理，将信息熵引入到空间对象群聚类算法中，替代了空间对象群相似性的比较。

定义 9(空间对象群聚类) $S = \{\boldsymbol{S}_1, \boldsymbol{S}_2, \text{L}\ , \boldsymbol{S}_m\}$ 是由空间数据集 D 构成的一个空间对象群的集合，用 CL 表示 S 的一个聚类，$CL = \{C_1, C_2, \text{L}\ , C_k\}$，$C_l$ 是第 l 个簇，$C_l \subseteq S$，$l \in [1,k]$，k 表示聚类个数，CL 使得目标函数 F 的值最小，F 按式(3)计算，其中，E_l 为 C_l 中所有空间对象群的信息熵的值，按式(4)计算：

$$F = \sum_{l=1}^{k} E_l \tag{3}$$

$$E_l = -\sum_{h=1}^{n}\sum_{t=1}^{T} p_{ht}^{l} \times \lg p_{ht}^{l} \tag{4}$$

3　基于信息熵的空间对象群聚类算法 ESOGC

$S = \{\boldsymbol{S}_1, \boldsymbol{S}_2, \text{L}\ , \boldsymbol{S}_m\}$ 是由空间数据集 D 构成的空间对象群的集合，空间对象群的聚类是将 S 中相似的空间对象群聚为一簇。基于信息熵的空间对象群聚类算法 ESOGC 可以分为 5 步：初始化，基于信息熵的蚁群聚类，聚类优化，聚类标记和聚类合并，最后输出聚类结果。

3.1　初始化

根据应用问题可以从空间数据集 D 中提取出主题对象 O_i；基于主题对象及相应的拓扑关系，可以将空间数据集划分成若干个空间对象群 $\boldsymbol{S}_i$。为了使空间对象群的表示维数一致，这里采用对空间数据集预处理的方法，如图 1 所示，方法如下：

(1)基于定义 2，将 D 按对象类型分成 n 层。

(2)将每层的对象按属性值分类，T_h 为第 h 层上的分类个数。

(3)基于定义 3，将每个空间对象群中的对象按层分组，B_{ih} 为 $\boldsymbol{S}_i$ 中的对象在 A_h 层上的分组，B_{ih} 可以为空；

(4)基于步骤(2)，将空间对象群各层上的对象分类，B_{ih}^{t} 为 B_{ih} 中基于 A_h 层分类的第 t 类对象的集合，B_{ih}^{t} 中包含的对象属性值均相似，B_{ih}^{t} 可以为空。

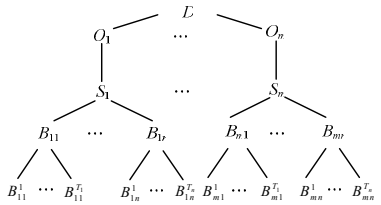

图 1　空间数据集预处理过程

初始化步骤中还包括一些参数的设置，如算法迭代次数 $N_{\text{iteration}}$、蚂蚁数 N_{ant}、蚂蚁初始负载等。

3.2　基于信息熵的蚁群聚类

基于信息熵的蚁群聚类是 ESOGC 算法的主要步骤：通过空间对象群主题对象邻域范围内信息熵的变化，指导蚂蚁捡起或放下空间对象群，实现空间对象群的聚类。步骤如下：

(1)For iter=1 to $N_{\text{iteration}}$

(2)　For q=1 to N_{ant}

(3)　If (ant_q.isload=false) then 蚂蚁 ant_q 随机拾起空间对象群 S_i，$i \in [1,m]$;

(4)　else 计算 S_i 主题对象邻域范围内 ant_q 未放下 S_i 前的信息熵的值 E_1;

计算 S_i 主题对象邻域范围内 ant_q 放下 S_i 后的信息熵的值 E_2;

(5)　If(E_1>= E_2||fail>N_{fail}) then ant_q 在当前位置放下 S_i;

(6)　else ant_q 负载 S_i 随机选择一个方向移动步长 d 到下一个位置，fail++;

(7)　End if

(8)　End if

(9)　End For

(10)End For

在迭代过程中，可以动态调整蚂蚁搜索半径 r 及步长 d 的值，这不仅提高了算法的效率，还降低了算法陷入局部最优的可能性。有限次迭代后，算法输出聚类结果。

3.3　聚类优化

基于信息熵的蚁群聚类结果是粗糙的，如图 2 所示，聚类过程中蚂蚁只考虑将相似的空间对象群堆积，相异的空间对象群分离，却没有考虑将不同的簇拉开，导致簇与簇之间划分很模糊。

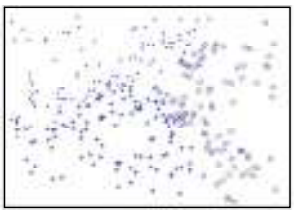

图 2　基于信息熵的蚁群聚类输出结果

由图 2 可以看出：对于簇内的空间对象群，其主题对象邻域范围内当前位置的信息熵总为最小，相反，对于簇边界上的空间对象群，其主题对象邻域范围内越是靠近同一簇，其信息熵则越小。基于这个原理，提出优化步骤如下：

(1) For i=1 to m　//m 为空间对象群总数

(2) S_i 的主题对象 O_i 的当前位置记为 p，蚂蚁拾起 S_i，蚂蚁失败次数 fail=0；

(3) 分别计算蚂蚁在位置 p 以及蚂蚁在 8 个方向上分别移动一步后 O_i 邻域范围内信息熵的值。取其中的最小值并记录其对应的空间位置 p’；

(4) If(fail<N_{fail}&&p≠p’) then 蚂蚁负载 S_i 步行到 p’位置，fail++，p=p’，转(3)；

(5) else 蚂蚁在 p’位置将 S_i 放下；

(6) End if

(7) End For

图 2 经过优化后的效果如图 3 所示，显然，不同的簇已完全拉开，同一簇则完全收拢。

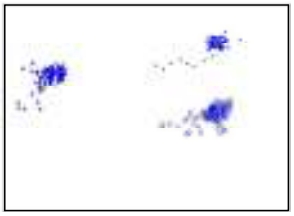

图 3 聚类优化输出结果

3.4 聚类标记

聚类标记就是将优化输出结果中距离较近的空间对象群标记为同一簇，距离较远的则标记为不同的簇。例如图 3 经过聚类标记后为 3 个簇。标记步骤如下：

(1)初始化集合 $N=\varnothing$，簇编号 $N_{\text{class}}=1$；

(2)随机从 S 中选择一个未被处理过的空间对象群 $\boldsymbol{S}_i$，$i\in[1,m]$；

(3)计算 $\boldsymbol{S}_i$ 的主题对象 O_i 的邻域 $N(O_i)$ 以及 $N=N\cup N(O_i)-N\cap N(O_i)$；

(4)对 N 中未被处理过的主题对象，重复执行步骤(3)，标记 N 中所有主题对象对应的空间对象群为第 N_{class} 簇；

(5) $N=\varnothing$，$N_{\text{class}}++$，转(2)，直到 S 中所有空间对象群均被处理过为止；

聚类标记将优化的蚁群聚类结果划分成若干个簇。

3.5 聚类合并

在蚁群算法聚类时，由于蚂蚁随机地捡起和放下操作，同一簇的空间对象群可能会被堆积成几个小簇，因此需要进行簇之间的合并。具体方法为：对于已标记的簇，分别计算该簇和其他簇合并前以及合并后的信息熵，若合并后信息熵不增大，则说明这 2 个簇可以合并，标记为同一簇。

4 实验结果分析

实验采用 2 个人工数据集 D_1、D_2 和 1 个实际数据集 D_3，每个人工数据集设计成 4 组空间对象群 C_1、C_2、C_3、C_4，每组包含 100 个空间对象群，所有对象分为 2 层。相关参数有：迭代次数 $N_{\text{iteration}}$=20 000，蚂蚁数 N_{ant}=5，步长 d=2，半径 r=5，允许失败次数 N_{fail}=100，这里 r 也是空间对象群主题对象邻域的距离阈值。在实验中，随着迭代次数增加 d 由 2 逐渐增大到 5，r 由 5 逐渐增大到 15。

实验 1 D_1 中空间对象群的特点：同组的空间对象群在同层上的对象按属性值均为同类，如 C_1 中满足 $|B_{ih}|=|B_{ih}^1|$；不同组的空间对象群在同层上的对象按属性值均为不同类；C_4 的第 2 层设计为空。

每组空间对象群按主题对象的坐标分布到二维空间中，如图 4 所示，符号“o”、“.”、“+”及“*”分别表示 4 组空间对象群主题对象所在的实际位置。对 D_1 执行 ESOGC 算法，同一簇的空间对象群被分成了多个小簇，经过聚类标记、合并后的结果如图 5 所示，这里的符号表示空间对象群经过蚂蚁搬运后主题对象到达的逻辑位置。由图 5 看出，聚类结果与 D_1 预期设计分组一致，说明 ESOGC 算法对 D_1 有效。

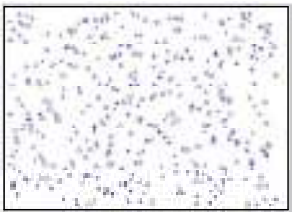

图 4 D_1 初始数据分布

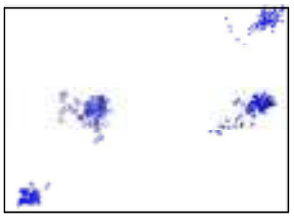

图 5 D_1 聚类结果

实验 2 D_2 中空间对象群的特点：同组的空间对象群在每一层上，对象属性值分类比例接近，例如，若 $\boldsymbol{S}_i$、$\boldsymbol{S}_j$ 为同一组，则 $\boldsymbol{S}_i$、$\boldsymbol{S}_j$ 满足 $|B_{ih}^t|/|B_{ih}|\approx|B_{jh}^t|/|B_{jh}|$；不同组的空间对象群在每一层上，对象属性值分类比例相差较大；C_4 的第 2 层设计为空。

D_2 中空间对象群的初始分布与图 4 相似。执行算法后，聚类结果与图 5 类似，与 D_2 预期设计分组一致，说明 ESOGC 算法对 D_2 有效。

实验 3 D_3 为文献[5]的数据集，该数据集包含了我国中 29 省(市、区)17 个森林样地数据，每一块森林样地包含了森林的树种、所属省(市、区)、经度、纬度、生物量、生产力等数据。每个省包含的森林样地构成一个空间对象群。基于森林样地中生物量和生产力属性值分布的差异，应用算法 ESOGC 聚类了我国森林生物生产力的地理空间格局，将我国各省划分为 9 类，分别是：(1)云南、四川、西藏、新疆；(2)青海；(3)湖南、江西；(4)福建、台湾、广西、广东；(5)海南；(6)江苏、浙江、安徽、湖北、贵州；(7)甘肃；(8)宁夏、河北、山西、山东、北京；(9)吉林、黑龙江、辽宁、内蒙古、陕西、河南。

文献[6]对我国森林生物生产力地理空间格局分析的结论是：“我国森林生物生产力随经度的增加呈线性递增，随纬度增加的递减速度受到经度的加快，即生物生产力的纬向递减率在我国西部地区较小，而在东部地区相对较大；而随海拔升高的递减速率又受到纬度变化的影响，即纬度越高，生物生产力随海拔增高的递减率越大。”算法 ESOGC 的聚类结果与此相符。文献[4]算法需要输入聚类数，而算法 ESOGC 在聚类过程中能自动确定聚类数，所以算法 ESOGC 的聚类结果比文献[4]更准确。

5 结束语

本文提出了基于信息熵的空间对象群聚类算法 ESOGC，解决了复杂空间数据集中具有拓扑关系的空间对象群的聚类问题，在对地理环境以及生物物种分布等的分析方面具有重要的应用价值。

(下转第 181 页)