文章编号: 1671-8860(2006)08-0679-04 文献标志码: A

从GIS数据库中挖掘空间离群点的一种高效算法

马荣华[1] 何增友[2]

(1 中国科学院南京地理与湖泊研究所, 南京市北京东路73号 210008)
(2 哈尔滨工业大学计算机系, 哈尔滨市西大直街92号, 150001)

摘 要: 根据GIS的空间特性, 借鉴已有的定义和概念, 定义了空间离群点是和在其非空间属性邻域内其他空间对象在空间位置上差异十分显著的空间对象, 并设计了SOD算法。实验结果验证了SOD算法的有效性和优越性, 给出了从GIS数据库中挖掘空间离群点的一般步骤。

关键词: GIS; 数据挖掘; 空间离群点; 最近邻
中图法分类号: P208

目前, 从GIS数据库中发现的基本知识类型主要包括普遍的几何知识、空间分布规律、空间关联规则、空间聚类规则、空间特征规则、空间区分规则、空间演变规则和面向对象的知识等[1-3], 而对GIS数据库中存在的小部分新颖的、与常规数据模式显著不同的新的数据模式则有所忽视。这些小部分异常数据构成了相对孤立的数据子集, 称为离群数据, 往往包含一些真实而又出乎意料的知识[4], 但在GIS关联规则发现和聚类分析等面向GIS数据库主体的数据挖掘任务中容易被作为噪声数据而被忽视。因此, 从GIS数据库中挖掘空间离群点具有十分重要的意义。Hawkins[5]揭示了离群点的本质; Shekhar等[6]提出了空间离群点的定义。随着研究的不断深入, 很多新的定义和概念被相继提出[7-10]。然而, 已有定义都是遵循Shekhar-Outlier的基本思想[6]。GIS数据库中的空间对象包括空间数据和属性数据, 具有空间属性和非空间属性, 另外还具有时间属性。一般而言, 一个空间对象的空间邻域是与该对象在空间属性上邻近的空间对象的集合, 而相互的比较则主要在非空间属性上进行。根据空间属性邻近关系的不同定义, 空间离群点主要包括两类, 一类是基于多维空间的, 一类是基于图连接的。前者以统计学为基础, 挖掘方法大致有4种[11]; 后者以图的连通性为基础, 将数据之间的关系映射到超图上, 用点与簇之间的连通度来评价, 从而有效地排除噪声对聚类结果的影响。根据GIS的空间特性, 本文从一个相反的角度定义空间离群点, 给出问题的定义以及离群点的检测算法, 最后给出应用该检测算法的GIS应用模式。

1 从新的角度看GIS空间离群点挖掘

在Shekhar等[6]的空间离群点定义中, 用空间属性定义邻域关系, 用非空间属性定义距离函数, 这种定义符合GIS的一般思维, 但在地理现象中经常会出现以相似的非空间属性为邻接的情况, 因此可以从一个相反的角度来定义和挖掘空间离群点, 即用非空间属性来定义邻域关系, 用空间属性来定义距离函数。换言之, 空间离群点是在和其非空间属性邻域内的其他空间对象在空间位置上差异十分显著的空间对象。容易看到, 新的定义在揭示空间离群点的本质和特性上, 与Shekhar等[6]提出的空间离群点的定义有相同的效果, 但具有如下的优势: ① 和Sun等[9]的方法类似, 新定义体现了空间离群点的局部特性; ② 基于新定义的算法容易实现, 不需要多维索引结构的支持, 对非空间属性值排序后可以很快地进行k-近邻搜索。

收稿日期: 2006-04-20。
项目来源: 国家自然科学基金资助项目(40301038)。

2 算 法

2.1 问题定义

令 $P \in R^{m+1}$, 对于 GIS 中任意空间对象 $p \in P$, 不妨设 p 有 m 个空间属性和 n 个非空间属性。令 $p, q \in P$, p 和 q 之间的空间距离表示为 $S\_D(kp, q)$, 即二者在 m 维空间上的 Euclidean 距离。同时, 用 $A\_D(p, q)$ 表示 p 和 q 在非空间属性上的距离。

此处, 将 Angiulli 等[12] 提出的基于距离的离群点的定义扩展到空间域。

定义 1 权重。令 k 为输入参数, P 为 GIS 空间数据集合, 且 p 是 P 中的一个对象, 则 p 的权重定义为 $\omega_k(p) = \sum_{i=1}^{k} S\_D(p, a^i(p))$。其中, a^i 表示 p 的第 i 个非空间属性近邻。

直观地讲, $a^i(p)$ 是在非空间属性上距离 p 最近的第 i 个对象, 恰好存在 $i-1$ 个对象 q 满足 $A\_D(q, p) \geqslant A\_D(q, a^i(q))$, 使得在非空间属性上距离 p 上存在 k 个点, 这 k 个点的集合构成了 p 的非空间属性邻域。定义 1 表明, 对象 p 的权重值为其非空间属性邻域内的对象与 p 的距离之和。显然, 该值越大, 表明 P 与其邻近的点的相异程度越大。

定义 2 空间离群点。P 为 GIS 空间数据集合, k 和 n 为参数, p 是 P 中的一个对象。如果恰好存在 $n-1$ 个对象 q, 且每个 q 满足 $\omega_k(q) \geqslant \omega_k(p)$, 则称 p 为第 n 个空间离群点, 用 $S\_Out_k^n$ 表示。同时, 用 $S\_O_k^n$ 表示 P 中 n 个离群点的集合。

因此, 给定参数 n(希望得到的空间离群点的个数)和 k(非空间属性邻域的大小)之后, 空间离群点检测的任务就是找到 n 个具有最大 ω_k 值的对象。

2.2 空间离群点检测算法(SOD)

基于上述定义, 本文给出一个新的空间离群点检测算法(图 1), 用 SOD 表示。分 3 个阶段进行。

1) 对空间对象按照非空间属性值进行排序。此阶段计算代价为 $O(N \cdot \log N)$, 其中 N 是 P 中空间对象的个数。

2) 确定每个对象的基于 k-近邻的非空间属性邻域, 并计算每个对象的权重。由于在上一阶段已经对空间对象按照非空间属性值进行了排序, 所以非空间属性 k-近邻邻域可以以 $O(k)$ 的代价完成。类似的, 每个对象权重的计算也可以以 $O(k)$ 的时间完成。于是, 这一阶段的时间复杂性为 $O(kN)$。

3) 对空间对象按照权重排序并输出 top-n 个离群点, 计算代价为 $O(N \cdot \log N)$。

因此, 整个 SOD 算法的时间复杂性为 $O(N \cdot \log N + kN)$。已有的算法中, 如文献[9], 即使在 R-Tree 等空间索引存在的情况下, 时间复杂性仍为 $O(N \cdot \log N + kN)$。因此, 从理论分析的角度, SOD 算法在执行时间上优于已有算法。

```
输入: P = {p1, p2, ..., pN}, 每个 pi ∈ P 是 m+1 维的空间对象; 最近邻
      的数目 k; 离群点的个数 n。
输出: n 个具有最大权值的空间离群点。

开始
/* ========第一阶段开始 ========== */
01   对非空间属性值进行排序
/* ========第一阶段结束 ========== */

/* ========第二阶段开始 ========== */
02   对于 P 中的每个对象 p
03     找到它在非空间属性上的 k 个最近邻。
04     计算 p 的权重, 即非空间属性邻域内的点与 p 的距离之和。
/* ========第二阶段结束 ========== */

/* ========第三阶段开始 ========== */
05   对上一阶段得到的权值按降序排序,
06   输出前 n 个具有最大权值的对象作为空间离群点 top-n。
/* ========第三阶段结束 ========== */
结束
```

图 1 SOD 算法

Fig. 1 Overview of SOD Algorithm

3 实 验

本文实验目的有两个: ① 比较 SOD 算法和 Shekhar 等[6]、Sun 等[9] 提出的算法在发现空间离群点上的效果; ② 验证 SOD 算法在大数据集中的可扩展性。对于第一个目标, 采用 Sun 等[9] 设计的数据集(图 2), 该数据集在 10×10 的网格上有 100 个数据对象, 属性数据服从高斯分布。分别用 SLZ 和 SLOM 表示 Shekhar 等[6] 和 Sun 等[9] 提出的算法, 表 1 给出了 SLZ、SLOM 和 SOD($k=5$) 分别发现的 5 个离群点的情况。

从表 1 可以看出, SOD 算法可以如 SLOM 算法那样, 有效地发现最为典型的局部空间离群点(8, 1), 此对象的异常情况从图 2 的数据分布中可以十分明显地看到, 然而 SLZ 算法却无法做到这一点。表 1 中所示的 SLZ、SLOM 以及 SOD 等 3 种算法所得的结果都可以直接从图 2 的数据分布中得到直观的解释。需要指出的是, 3 种算法得

到的离群点差异较大, 其原因在于该数据集的离群点较多, 所以不同方法由于各自的特点不同, 都能得到有意义但有一定差异的离群点。

为了测试 SOD 算法的可扩展性, 随机产生了 1 个具有 10 万个在 10 000 × 10 网格上的空间对象(属性数据同样服从高斯分布)。同时, 分别令 k= 3、5、7、9, 并将对象的数目从 20 000 变化到 100 000, 进而观察执行时间的变化情况。算法用 JAVA 实现, 测试结果表明, SOD 算法具有良好的可扩展性, 适合在大规模的空间数据库上进行离群点检测, 见图 3。

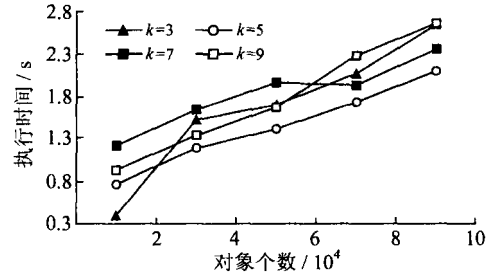

图 3　可扩展性测试

Fig. 3　Scalability Test

	0	1	2	3	4	5	6	7	8	9
0	-16	-9	-16	4	8	25	-2	20	-11	9
1	-3	-1	9	-12	1	1	-1	-2	-4	-2
2	14	1	11	2	-13	15	4	3	11	19
3	-10	16	-11	-2	-10	-11	-17	4	8	-15
4	-5	20	-11	4	-5	8	6	6	-2	-1
5	15	10	-9	7	12	-9	-18	16	8	-6
6	0	0	0	0	-21	-5	12	-15	-5	11
7	0	0	0	0	5	6	1	1	-9	3
8	0	8	0	0	-9	-8	-1	-2	9	5
9	0	0	0	0	19	-1	-2	-7	-3	-12

图 2　一个简化的空间数据集

Fig. 2　Synthetic Spatial Dataset

表 **1**　不同算法分别发现的 **5** 个离群点的情况

Tab. 1　Outliers Found by Different Methods on Same Dataset

位置 (SLZ)	$g(x)$ 值 (SLZ)	位置 (SLOM)	SLOM	位置 (SOD)	权重
(0, 7)	24.0	(0,5)	0.427 7	(9, 9)	44.1
(0, 5)	23.6	(2,5)	0.247 9	(1, 0)	40.9
(6, 4)	- 23.0	(0,7)	0.206 1	(0, 1)	38.9
(9, 4)	22.6	(8,1)	0.173 9	(1, 1)	38.2
(3, 9)	- 22.0	(3,9)	0.172 7	(8, 1)	37.4

4　结论与讨论

本文从新的角度定义了空间离群点, 新的定义具有如下的优势: ① 基于新定义的 SOD 算法不需要多维索引结构的支持, 执行时间优于已有算法; ② 新定义体现了空间离群点的局部特性。

SOD 算法可以直接应用于 GIS 数据挖掘中, 借鉴文献[3]中从 GIS 数据库中挖掘空间关联规则的步骤, 从 GIS 数据库中挖掘空间离群点的一般步骤可以概括为以下方面(见图 4)。

1) 提取图层, 建立相关数据库, 确定必要的参数。即从 GIS 数据库中提取与任务相关的图层或数据集, 存入另外的数据库中(称为相关数据库), 作为 SOD 算法中空间对象的数据集 P, 同时根据任务要求和经验, 确定参数 k 和 n。

2) 获取空间谓词[3, 13], 连接属性数据。通过已有的标准谓词库, 调用谓词函数, 进行空间关系计算, 存入另外的数据库中, 为 SOD 算法中权重值的计算作准备, 同时可以用于其他规则模式(如空间关联规则)的数据挖掘, 然后链接与任务相关的属性数据, 另存入关系数据库, 为 SOD 算法中 k- 近邻搜索作准备。

3) 执行 SOD 算法。将发现的离群点(包括空间数据和属性数据)存入离群点数据库。

4) 检验验证。如果不符合实际情况, 则调整参数 k 和 n, 重复前面的步骤; 否则, 把去除离群点后的数据存入结果数据库。

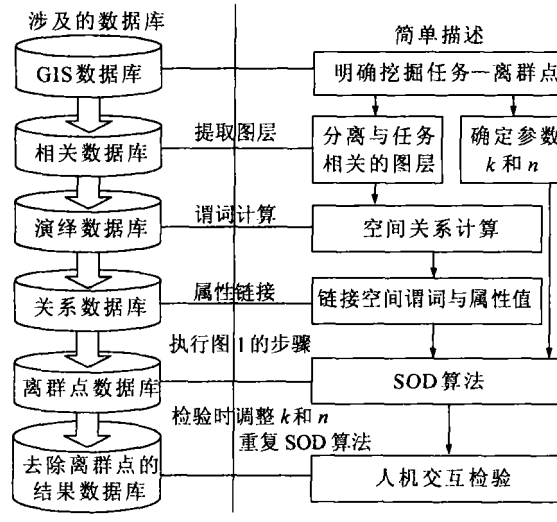

图 4　从 GIS 数据库中挖掘空间离群点的一般步骤

Fig. 4　A General Procedure to Mine Spatial Outliers from GIS

参 考 文 献

[1] 李德仁, 程涛. 从 GIS 数据库中发现知识[J]. 测绘学报, 1995, 24(1): 37-43

[2] 邸凯昌, 李德仁, 李德毅. 空间数据发掘和知识发现的框架[J]. 武汉测绘科技大学学报, 1997, 22(4): 328-332

[3] 马荣华, 马晓冬, 蒲英霞. 从 GIS 数据库中挖掘空间关联规则[J]. 遥感学报, 2005, 9(6): 733-741

[4] Knorr E M, Ng R T. A Unified Approach for Mining Outliers: Properties and Computation[C]. International Conference on Knowledge Discovery and Data Mining, Newport Beach, California, 1997

[5] Hawkins D. Identification of Outliers[M]. London: Chapman and Hall, Reading, 1980

[6] Shekhar S, Lu Changtien, Zhang Pusheng. A Unified Approach to Detecting Spatial Outliers[J]. GeoInformatica, 2003, 7(2): 139-166

[7] Lu Changtien, Chen Dechang, Kou Yufeng. Algorithms for Spatial Outlier Detection[C]. The 3rd IEEE International Conference on Data Mining, Melbourne, Florida, 2003

[8] Lu Changtien, Chen Dechang, Kou Yufeng. Detecting Spatial Outliers with Multiple Attributes[C]. 15th IEEE International Conference on Tools with Artificial Intelligence, Sacramento, California, 2003

[9] Sun Pei, Chawla S. On Local Spatial Outliers[C]. The 4th IEEE International Conference on Data Mining, Brighton, 2004

[10] Hu Tianming, Sung S Y. A Trimmed Mean Approach to Finding Spatial Outliers[J]. Intelligent Data Analysis, 2004(8): 79-95

[11] 魏藜, 宫学庆, 钱卫宁, 等. 高维空间中的离群点发现[J]. 软件学报, 2002, 13(2): 280-290

[12] Angiulli F, Pizzuti C. Outlier Mining in Large High-Dimensional Data Sets[J]. IEEE Trans on Knowledge and Data Engineering, 2005, 17(2): 203-215

[13] Ma Ronghua, Ma Xiaodong. Cognitive Logic Representation of Spatial Association Rules of Knowledge Discovery from GIS Database[C]. International Symposium on Spatial-Temporal Modeling, Spatial Reasoning, Spatial Analysis, Data Mining and Data Fusion, Beijing, 2005

第一作者简介: 马荣华, 博士, 副研究员。现主要从事GIS数据挖掘与认知理论以及湖泊水质遥感研究。

E-mail: rhma@niglas.ac.cn

Fast Mining of Spatial Outliers from GIS Database

MA Ronghua[1] *HE Zengyou*[2]

(1 Nanjing Institute of Geography and Limnology, CAS, 73 East Beijing Road, Nanjing 210008, China)

(2 Department of Computer Science and Engineering, Ha'erbin Institute of Technology, 92 Xidazhi Street, Ha'erbin 150001, China)

Abstract: On the basis of the spatial characteristics of GIS, an alternative viewpoint for defining and discovering spatial outliers from GIS is proposed, in which the spatial location of a spatial outlier is significantly far from other spatial objects in its neighborhood determined by non-spatial attribute. Then the SOD algorithm is proposed and analyzed in detail. Experimental results on synthetic datasets demonstrate that the proposed approach can effectively and efficiently identify spatial outliers in large spatial data sets. Finally, the authors give the general procedure to mine the spatial outliers from GIS database.

Key words: GIS; data mining; spatial outlier; nearest neighbor

About the first author: MA Ronghua, Ph. D, associate research fellow. He is concentrated on the research in data mining from GIS and cognitive theory, and remote sensing of water quality of inland lakes.

E-mail: rhma@niglas.ac.cn