基于空间相邻关系的 GML点对象离群检测算法

陈佳春[1,2], 吉根林[1,2]

(1. 南京师范大学 数学与计算机科学学院, 江苏 南京 210097;
2. 南京师范大学 虚拟地理环境教育部重点实验室, 江苏 南京 210097)

[摘要] 提出了一种基于空间相邻关系的点对象离群检测算法 SAOD(Space Adjacent Relations Based GML Point Outlier Detection Algorithm). 利用空间相邻关系作为空间点对象的相似度度量准则, 得到相似度矩阵, 从而挖掘 GML中的离群点对象. 实验结果表明, SAOD算法能有效地检测 GML中的离群点对象并且具有较高的效率.

[关键词] 离群点检测, 空间相邻, GML数据挖掘
[中图分类号] TP311 [文献标识码] A [文章编号] 1672-1292(2009)01-0061-03

An Algorithm for GML Point Outlier Detection Based on Space Adjacent Relations

Chen Jiachun[1,2], JiGenlin[1,2]

(1. School of Mathematics and Computer Science, Nanjing Normal University, Nanjing 210097, China;
2. Key Laboratory of Virtual Geographic Environment, Ministry of Education, Nanjing Normal University, Nanjing 210097, China)

Abstract At present algorithm for GML outlier detection has seldom been researched. Algorithm SAOD for GML point outlier detection based on space adjacent relations is proposed in this paper. In this algorithm, the space adjacent relations between spatial points are considered as the similarity measurement and similarity matrix is computed. The expected outliers can be obtained from the matrix. The results of experiments show that algorithm SAOD is effective and efficient.

Key words outlier detection; space adjacent relations; GML data mining

离群点(outlier)检测是数据挖掘的基本任务之一. 国内外研究者已经提出了许多离群点检测算法, 例如, 基于统计学的离群点检测算法[1]假设一个分布或概率模型, 然后根据模型采用不一致性检验来确定离群点. 该方法的一个主要缺点是绝大多数检验是针对单个属性的, 而且统计学方法不能确保所有的离群点被发现. DBOD算法[2]是通过数据点或对象之间的距离来检测离群点. 该算法与基于统计的方法相比拓展了多个标准分布的不一致性检测的思想, 但此方法要求用户设置距离参数, 可能涉及多次的试探和错误. 文献[3]提出了一种基于密度的离群检测算法, 但在决定离群点方面, 该算法更强调对象的局部性. 文献[4]提出了一种基于空间物体研究多个非空间属性的空间离群点的检测方法, 但它仅能对点进行离群检测, 并未涉及空间相邻的拓扑关系.

GML(Geography Markup Language)是一种 Internet环境下的空间信息编码方式, 用于数据传输和存储. 随着 Internet的应用, GML数据不断增多, 研究面向 GML数据的离群挖掘算法具有理论意义和应用价值. 目前基于拓扑关系的 GML空间离群检测算法尚无文献报道, 为此本文提出基于空间相邻关系的 GML离群点检测算法 SAOD. 该算法对 GML文档中的点空间对象进行相邻关系分析, 并将之作为空间对象相似性度量准则, 从而挖掘 GML文档中的离群空间对象. 实验结果表明, 本文提出的基于空间相邻关系的 GML离群点检测算法 SAOD是有效的, 具有较高的执行效率.

收稿日期: 2008-12-10
基金项目: 国家自然科学基金(40771163, 40871176)资助项目.
通讯联系人: 吉根林, 教授, 博士生导师, 研究方向: 数据挖掘技术及应用技术、XML技术. E-mail: glj@njnu.edu.cn

1　相关概念

定义 **1**　设一个空间点对象 $D_1(x_1, y_1)$, 点对象 $D_2(x_2, y_2)$ 与 D_1 满足空间相邻关系当且仅当 D_1 与 D_2 的距离 d 在一定的阈值范围之内, 也即:

$$d(D_1, D_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \leqslant M, \tag{1}$$

式中, M 为给定的最短距离阈值.

定义 **2**　若空间点对象 $D_1, D_2, \cdots, D_n$ 均与对象 D 满足空间相邻关系, 则称 $\{D_1, D_2, \cdots, D_n\}$ 为 D 的邻域 $N(D)$.

定义 **3**　设两个点空间对象 D_1、D_2, D_1, D_2 的邻域分别为 $N(D_1) = \{a_1, a_2, \cdots a_m\}$, $N(D_2) = \{b_1, b_2, \cdots b_n\}$. 那么 D_1、D_2 基于空间相邻关系的相似度为

$$\mathrm{Sim}(D_1, D_2) = \frac{|N(D_1) \cap N(D_2)|}{|N(D_1) \cup N(D_2)|}, \tag{2}$$

式中, $|N(D_1) \cap N(D_2)|$ 为 D_1 与 D_2 相同的相邻对象的个数, $|N(D_1) \cup N(D_2)|$ 为 D_1 与 D_2 包含的相邻对象总数目.

定义 **4**　设 minsim 为最小相似度阈值, $\mathrm{Sim}(D_i, D_j)$ 为空间点对象 D_i 与 D_j 之间的相似度, $SA[i]$ 为空间点对象 D_i 与其他点对象之间的平均相似度, 若 $SA[i] < \mathrm{minsim}$, 则称 D_i 为空间离群点对象.

2　算法 *SAOD*

2.1　算法思想

算法 SAOD使用 GML文档中点空间对象之间的相邻关系作为空间对象相似性的度量准则, 对空间点对象进行离群挖掘, 算法步骤如下:

(1) 解析 GML文档, 获取 GML文档中的空间对象集合;

(2) 对空间点对象进行相邻关系计算, 获得点空间对象的邻域;

(3) 根据空间点对象的相邻关系, 计算点空间对象两两相似度得到相似度矩阵;

(4) 对相似度矩阵进行计算与分析, 得到空间离群点.

2.2　相似度矩阵的计算

现举例说明: 设有 4个同类的空间点对象 D_1, D_2, D_3, D_4, 通过空间对象的相邻关系计算, 得出空间相邻关系表, 如表 1所示.

表 **1**　空间相邻关系表

Table 1　Space adjacent relations table

空间点对象 D_i	空间相邻点对象 $N(D_i)$
D_1	a, b, c, d, f
D_2	a, d, g
D_3	c, d, m
D_4	m, p

对表 1经过计算可得: $\mathrm{Sim}(D_1, D_2) = \frac{|N(D_1) \cap N(D_2)|}{|N(D_1) \cup N(D_2)|} = 0.33$, $\mathrm{Sim}(D_1, D_3) = \frac{|N(D_1) \cap N(D_3)|}{|N(D_1) \cup N(D_3)|} = 0.33$, 同理可以计算出所有两两点对象之间的相似度, 从而得到相似度矩阵 $\boldsymbol{S}$:

$$\boldsymbol{S} = \begin{bmatrix} 1.00 & 0.33 & 0.33 & 0 \\ 0.33 & 1.00 & 0.20 & 0 \\ 0.33 & 0.20 & 1.00 & 0.25 \\ 0 & 0 & 0.25 & 1.00 \end{bmatrix}.$$

通过相似度矩阵 $\boldsymbol{S}$, 可以计算出每个对象 D_i 与其他对象 D_j 的平均相似度, 计算方法如式(3)所示. 每个对象 D_i 与其他对象 D_j 的平均相似度的计算不应考虑 D_i 自身与自身的相似度, 而是要计算 D_i 与其他 $(n-1)$ 个对象的平均相似度. 对象自身与自身的相似度为 1.

$$SA[i] = \left(\sum_{j=1}^{n} \boldsymbol{S}[i][j] - 1 \right) / (n-1). \tag{3}$$

对于上述相似度矩阵 $\boldsymbol{S}$, 其 $SA[1] = 0.22$, $SA[2] = 0.18$, $SA[3] = 0.26$, $SA[4] = 0.08$, 设 minsim = 0.1, 由于 $SA[4] < \mathrm{minsim}$, 因此认为 D_4 为离群点对象.

2.3　算法描述

基于空间相邻拓扑关系的 GML离群点对象检测的算法 SAOD描述如下:

输入: GML文档 Doc; 最小相似度阈值 minsim;
输出: GML空间离群点集合 O;
步骤:
(1) P = interpret(Doc); //解析 GML文档得到空间点对象集合 P
(2) $O = \emptyset$; //初始化 O 为空集
(3) for each D in P do
$N(D)$ = compute_adj(P); //得到 D的空间相邻关系
(4) for(i= 1; $i \leqslant |P|$; i++) do
for(j= 1; $j \leqslant |P|$; j++) do
$S[i][j] = \mathrm{sim}(N(D_i), N(D_j))$;
(5) for(i= 1; $i \leqslant |P|$; i++) do
$SA[i] = \left(\sum_{j=1}^{|P|} S[i][j] - 1 \right) / (n-1)$;
(6) sort(SA); //对数组 SA 进行排序
(7) for(i= 1; $i \leqslant |P|$; i++) do
if($SA[i] <$ minsim) $O = O \cup D_i$;

2.4 算法的时间复杂度分析

设 GML文档中空间对象的个数为 n, 解析 GML文档的时间复杂度为 $O(n)$, 计算空间相邻关系和相似度矩阵的时间复杂度均为 $O(n^2)$, 计算对象之间平均相似度的时间复杂度为 $O(n)$, 用快速排序实现相似度排序的时间复杂度为 $O(n \log n)$, 最后判断是否对象是离群点, 其时间复杂度为 $O(n)$, 因此整个算法的时间复杂度为 $O(n^2)$.

3 结果与分析

为验证本文提出的基于空间相邻关系的 GML离群点检测算法 SAOD的有效性, 我们在 Intel Pentium Ⅳ 2.93 GHz, 内存 512 M 的 PC机上, 将该算法利用 VC++加以实现, 实验中 GML文档数据由程序自动生成. 由于目前基于 GML的离群点检测算法尚未见文献报道, 因此算法 SAOD的性能无法与同类算法进行比较. 算法 SAOD 的执行效率如图 1所示. 对于程序自动生成 GML 文档, 给定 minsim = 0.01, 当 GML文档的大小分别为 10 k, 20 k, 50 k, 80 k, 100 k时, 检测到的离群点对象的个数对应为 0, 2, 5, 8, 14. 实验表明, 算法 SAOD 能有效检测基于相邻空间关系的 GML数据中离群点对象.

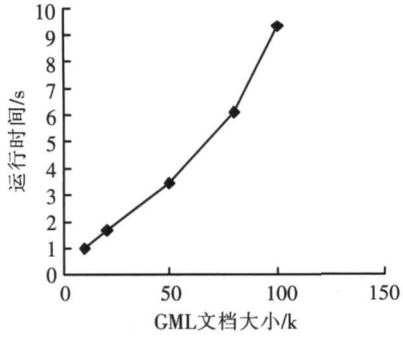

图 1 SAOD 算法的执行时间
Fig.1 The executing time of SAOD

4 结 语

本文对基于空间相邻关系的 GML点对象离群检测算法进行研究, 提出了基于空间相邻关系的 GML点对象离群检测算法 SAOD. 该算法通过引入空间相邻关系, 对 GML文档中点对象数据进行离群检测. 实验表明该算法是有效的. 除了点对象相邻关系离群检测以外, 还有点、线、面等空间拓扑关系的离群检测有待进一步研究.

[参考文献] (References)

[1] Barnett V, Lewis T. Outliers in Statistical Data[M]. New York: John Wiley & Sons, 1994: 194-223.
[2] Knorr E M, Ng R T. Algorithms for mining distance-based outliers in large datasets[C]//Proc of 1998 International Conference Very Large Data Base (VLDB98). New York: VLDB Endowment, 1998: 392-403.
[3] Breunig M M, Krueger H P, Ng R, et al. LOF: identifying density-based local outliers[C]//Proc of ACM SIGMOD2000 International Conference on Management of Data. Dallas, Texas: ACM Press, 2000: 93-104.
[4] Knorr E M, Ng R T. A unified notion of outliers: properties and computation[C]//Proc the 3rd International Conference on Knowledge Discovery and Data Mining. California: IEEE Press, 1997: 219-222.

[责任编辑: 严海琳]