空间离群点的检测算法

贾瑞玉,钱光超,张 然,李龙澍
(安徽大学 计算机科学与技术学院,安徽 合肥 230039)

摘 要:空间离群点是指与其邻居具有明显区别的属性值的空间对象。已有的空间离散点检测算法一个主要的缺陷就是这些方法导致一些真正的离群点被忽略而把一些非离群点当成了空间离群点。提出了一种迭代算法,该算法通过多次迭代检测离群点,取得较好效果。实验表明该算法具有较好的实用性。

关键词:空间离群点检测;地理信息系统;迭代算法
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2008)05-0028-03

A Spatial Outlier Detection Algorithm

JIA Rui-yu,QIAN Guang-chao,ZHANG Ran,LI Long-shu
(School of Computer Science and Technology, Anhui University, Hefei 230039, China)

Abstract: A spatial outlier is a spatially referenced object whose non-spatial attribute values are significantly different from the values of its neighborhood. The identification of outliers can lead to the discovery of useful knowledge and detecting spatial outliers is useful in many applications of geographic information systems. One major drawback of the existing detection approaches is that their application will lead to some true spatial outliers being ignored and some false spatial outliers being identified. In this paper, propose an iterative algorithm that detects spatial outliers by multi-iterations. This algorithm gets a good performance. Experiments show that this algorithm has a good practicality.

Key words: spatial outlier detection; GIS; iterative algorithms

0 引 言

离群点查找是旨在发现偏离常规模式的小部分异常模式的过程。在某些领域里,研究离群点的异常行为有助于发现特别有价值的知识[1]。在空间数据库中,由于对象存在空间关系,因此空间对象的离群点挖掘方法与传统的面向统计数据的离群点挖掘方法有区别[2],应考虑空间关系,这就导致了开展空间离群点的研究。

至今,空间离群点还没有统一的模型或定义,由于其特殊的应用价值,因此一些学者认为它是由某种特有的机制产生的。Shekhar 等[2]认为空间离群点是指空间数据中与其他对象不一致的对象,即空间邻域中非空间属性与其他对象明显不同的空间对象,而且,它是局部不稳定的,即使对总体来说并不异常,但对邻近的其他点却具有极端的值。空间离群点查找在交通、地质、公共安全、公共健康、气象以及基于空间位置的服务等方面有着广泛的应用[3]。

已有的查找离群点研究大多没有区分地理空间维与属性空间维。其中地理空间维是指空间属性,包括点、线、面等几何属性与拓扑属性;属性空间维是指非空间属性,包括长度、归属、地名与各种非空间度量值等等。这两种属性不区分的模型不能反映离群点偏离的实际情况,从而可能发现不正确的离群点。

现有的空间离群点检测方法一个主要的缺陷就是这些方法导致一些真正的离群点被忽略而把一些非离群点当成了空间离群点,而且大多不是基于局部的,它可能找到全局离群点,却可能丢失局部离群点。因为全局离群点容易识别,甚至是已知,所以对于知识发现来说,基于局部的离群点往往比全局离群点更有意义。鉴于此,文中提出了一种迭代算法,该算法通过多次迭代检测离群点,并在迭代过程中对离群点的属性值进行修正,可以提高检测结果的正确性,并能检测局部离群点。

1 问题描述

给定 p 维空间对象点集 $X=\{x_1,x_2,\cdots,x_n\}$(p

收稿日期:2007-08-20
基金项目:安徽省教育科研项目(2005kj056)
作者简介:贾瑞玉(1965-),女,河南淅川人,副教授,研究方向为计算机图形学、数据挖掘、人工智能。

$\geqslant 1)$,X 中的一个空间对象点 x 由 p 个变量属性值 $(y_1,y_2,\cdots,y_p)$,记为 $\boldsymbol{Y}(y_1,y_2,\cdots,y_p)^{\mathrm{T}}$,T 是矩阵转置运算符。属性函数 f 定义为 $X\rightarrow R^p$(R^p 是 p 维欧氏空间)映射,使得每个空间点 x 的函数值 $f(x)=y$ 属性向量[4] 记为:

$$\boldsymbol{Y}_i=f(x_i)=(f_1(x_i),f_2(x_i),\cdots,f_p(x_i))^{\mathrm{T}}=(y_{i1},y_{i2},\cdots,y_{in}),i=1,2,\cdots,n。$$

给定正整数 k,记 $\mathrm{NN}_k(x_i)$ 为点 x_i 的 k 个最近邻居集($k\geqslant 1,i=1,2,\cdots,n$)。邻居函数 g 定义为 $X\rightarrow R^p$ 映射,$g(x)$ 的第 j 个分量记为 $g_j(x)$,$g_j(x)$ 是 $\mathrm{NN}_k(x)$ 中所有空间点的属性值 y_j 的一个概括统计,例如 $g(x)$ 可以是点 x 的 $\mathrm{NN}_k(x)$ 中所有空间点的属性值的均值。

为检测空间离群点,比较与 x 相邻的 x 中 y 的所有分量。比较函数 h 是关于 f 和 g 的函数,其定义域是 X,值域是 $R^r(r\leqslant p)$。例如,$h=f-g$ 代表了 $X\rightarrow R^p(r=p)$ 的映射,$h_1=f_1/g_1$ 代表 $X\rightarrow R(r=1)$ 的映射。记 $h(x_i)$ 为 $h_i,i=1,2,\cdots,n$。

给定属性函数 f,邻居函数 g,比较函数 h,如果 h_i 是集合 $\{h_1,h_2,\cdots,h_n\}$ 中的极值,则对应的 x_i 是空间离群点。由此可以看出空间离群点的确定依赖于函数 g 和 h 的选择。

空间孤立点检测问题一般化形式如下:

给定一个空间点集 $X=\{x_1,x_2,\cdots,x_n\}$;其中,邻居关系为 $\mathrm{NN}_k(x_1)$,$\mathrm{NN}_k(x_2),\cdots,\mathrm{NN}_k(x_n)$;属性函数 $f:X\rightarrow R^p$,邻居函数 $g:X\rightarrow R^p$,比较函数 $h:X\rightarrow R^r(r\leqslant p)$,空间孤立点数目 m,阈值 θ。据此设计算法来检测空间孤立点。

2　空间离群点检测算法

假定所有的 $k(x_i)$ 都是一固定的整数 k(空间点 x_i 的邻居数设定为 $k(x_i)$,也可根据需要动态设定)。根据第1节的分析可知空间离群点检测算法依赖于邻居函数 g 和比较函数 h 的选择,即 g 和 h 的选择决定了算法的效果。在文中的算法中,点 x 的邻居函数 g 取 x 的 k 个最近邻居的属性值的平均值(这样可以减少具有较大或较小属性值的邻居点的影响)。比较函数 $h(x)=f(x)-g(x)$,于是得到集合 $\{h_1,h_2,\cdots,h_n\}$(n 是对象数目)。对于某个 h_i 是集合 $\{h_1,h_2,\cdots,h_n\}$ 中的极值,则其对应的 x_i 是个候选的离群点。令 μ、σ 分别是 $\{h_1,h_2,\cdots,h_n\}$ 的样本均差和样本标准差,则 h_i 标准化之后的值为 $h'_i=\dfrac{h_i-\mu}{\sigma}$,得到 $\{h_1,h_2,\cdots,h_n\}$ 标准化之后的数据集 $\{h'_1,h'_2,\cdots,h'_n\}$。因此,当且仅当 h'_i 是标准化数据集 $\{h'_1,h'_2,\cdots,h'_n\}$ 中的极值时相应的点 x_i 也是原始数据集 $X=\{x_1,x_2,\cdots,x_n\}$ 中的极值,相应地,x_i 也是一个可能的空间离群点,如果 $|h'_i|$ 足够大的话(通过阈值 θ 来判断)。

算法详细说明如下:

(1) 计算数据集 $\{x_1,x_2,\cdots,x_n\}$ 中每个空间点 x_i 的 k - 最近邻居集合 $\mathrm{NN}_k(x_i)$,邻居函数 $g(x_i)=\dfrac{1}{k}\sum_{x\in \mathrm{NN}_k(x_i)}f(x)$,比较函数 $h_i=h(x_i)=f(x_i)-g(x_i)$。

(2) 计算集合 $\{h_1,h_2,\cdots,h_n\}$ 的样本均差 μ 和样本标准差 σ,对集合 $\{h_1,h_2,\cdots,h_n\}$ 进行标准化得到集合 $\{h'_1,h'_2,\cdots,h'_n\}$,令 $y_i=|h'_i|=\left|\dfrac{h_i-\mu}{\sigma}\right|(i=1,2,\cdots,n)$,得到集合 $\{y_1,y_2,\cdots,y_n\}$,y_q 是集合 $\{y_1,y_2,\cdots,y_n\}$ 中的最大值。对于给定阈值 θ,如果 $y_q\geqslant\theta$,则 y_q 对应的 x_q 是一个空间孤立点。

(3) 令 $f(x_q)=g(x_q)$,即把 x_q 修正为 x_q 的 k 个最近邻居的属性值的均值,这样可以降低 x_q 对下一步操作的影响。对于每一个 x_i,如果其 k - 最近邻居集合 $\mathrm{NN}_k(x_i)$ 中包含 x_q,则重新计算其 $g(x_i)$ 和 h_i。

(4) 重新计算 $\{h_1,h_2,\cdots,h_n\}$ 的样本均差 μ 和样本标准差 σ,重新标准化 $\{h_1,h_2,\cdots,h_n\}$ 得到 $\{h'_1,h'_2,\cdots,h'_n\}$,重新计算 $y_i=|h'_i|=\left|\dfrac{h_i-\mu}{\sigma}\right|,i=1,2,\cdots,n$。

(5) 重复步骤(2)~(4)直到阈值条件无法满足或者空间孤立点的数目已达到 m。

阈值 θ 一般选定为2或者3。这是因为在该算法中,如果属性函数 f 是正态分布的话那么比较函数 h 一般也是正态分布[2]。

从上述算法中可以看到,一旦检测到一个空间离群点,将进行一系列更新操作,包括用空间离群点的 k 个邻居点的属性值的均值来替换离群点的属性值以及一系列的更新计算。这些更新操作的目的是为了避免将与空间离散点较近的空间点误认为是离散点。

3　实验演示

通过实验比较几个算法的挖掘结果。

对于图1中所示意的数据(其中,S1、S2、S3 是三个真正的空间离群点)分别采用文中的算法、文献[5]中的 Scatterplot 算法和文献[6]中的 Moran Scatterplot 算法进行对比实验。参数设置:$k=m=3$。

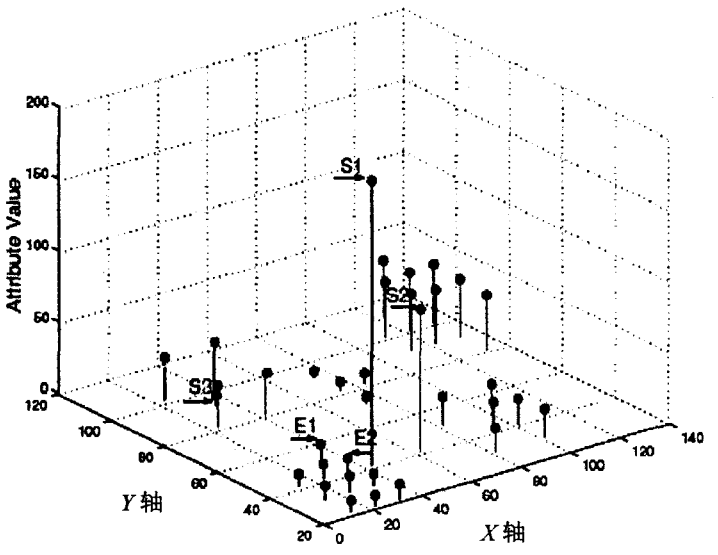

图 1 空间数据集

图中，空间对象定位在 $X-Y$ 平面，竖轴的高度表示每个对象的属性值。

表 1 给出了实验的比较结果，可以看出，文中的算法准确地检测到三个真正的空间离群点 S1、S2、S3，而 Scatterplot 算法和 Moran Scatterplot 算法则均错误地把 E1 和 E2 当成了空间离群点。实验的结果表明文中提出的算法可以成功地挖掘出空间离群点而同时避免了将非空间离群点错误地当成了空间离群点。

表 1 实验结果

离群点	算法		
	Scatterplot	Moran Scatterplot	文中算法
1	E1	S1	S1
2	E2	E1	S2
3	S2	E2	S3

4 结束语

文中提出了一种基于迭代的空间离群点检测算法，该算法通过多次迭代检测离群点，并在迭代过程中对离群点的属性值进行修正，可以提高检测结果的正确性，并能检测局部离群点。实验验证了该算法的有效性。

参考文献：

[1] Tan Pang-Ning, Steinbach M, Kumar V. Introduction to Data Mining[M]. 北京：人民邮电出版社，2006.

[2] Shekhar S, Lu C T, Zhang P. A Unified Approach to Spatial Outlier Detection GeoInformalica [J]. International Journal on Advances of Computer Science for Geographic Information System, 2003, 7(2): 139-166.

[3] 王占全. 基于 GIS 空间数据挖掘若干关键技术的研究[D]. 杭州：浙江大学，2005.

[4] Lu Chang-Tien, Chen Dechang, Kou Yufeng. Detecting Spatial Outliers with Multiple Attributes[C]// Proceedings of the 15th International Conference on Tools with Artificial Intelligence. [s.l.]: [s.n.], 2003.

[5] Haining R. Spatial data Analysis in the Social and Environmental Sciences[M]. Cambridge: Cambridge University Press, 1993.

[6] Luc A. Local Indicators of Spatial Association: LISA[J]. Geographical Analysis, 1995, 27(2): 93-115.

(上接第 27 页)

3 结束语

基于三维空间的平行边有向图构造算法和概念格的三维重构机制有效地解决了概念格在三维空间中的节点定位布局和线段交叉问题，使格结构更加清晰易懂。而且采用这种重构方法构造出的概念格具有特别良好的结构，增强了概念格的表达能力，为知识处理提供了良好的数据基础。

但是，采用这种方法进行构造，由于系统对于复杂的概念格需要进行形式背景的拆分并构造子格，系统的时间复杂度和空间复杂度将会成倍增加，因此寻找一种更优化的算法将是今后进一步的研究目标；同时由于子格的合并是一个非常复杂的过程，如何简化合并过程并得到精确的结果也将是今后研究的重点。

参考文献：

[1] Wille R. Lattices in data analysis: How to draw them with a computer[M]. Berlin: Berlinpringer, 1993: 47-158.

[2] Krohn U, Davies N J, Weeks R. Concept lattices for knowledge management[J]. BT Technol J, 1999, 17(4): 106-114.

[3] Cole R. Automated Layout of Concept Lattice Using Layer Diagrams and Additive Diagrams[C]//Concept Lattices: Second International Conference on Formal Concept Analysis, ICFCA 2004. Sydney: Gri_th University, 2004: 31-42.

[4] Sugiyama K, Tagawa S, Toda M. Methods for visual understanding of hierarchical system structures[J]. IEEE Transactions on Systems, Man and Cybernetics, 1981, 11(2): 109-125.

[5] 谢 润，李海霞，马 骏，等. 概念格的分层及逐层建格法[J]. 西南交通大学学报，2005，40(6)：837-841.

[6] 马 骏，沈夏炯，刘宗田. 基于三维空间的概念格自动布局[J]. 计算机科学，2006，33(5)：244-246.

[7] Ganter B, Wille R. Formal Concept Analysis: Mathematical Foundations[M]. [s.l.]: Springer Verlag, 1999.