基于裁边策略的空间离群区域检测算法*

陈 涛, 万家强, 朱庆生
(重庆大学 计算机学院, 重庆 400044)

摘 要: 为了检测空间数据集中存在的离群区域,提出一种基于裁边策略的空间离群区域检测算法。首先利用 Delaunay 三角网格确定空间邻接关系,根据非空间属性描述邻接节点间的差异性;然后反复裁去最大权边,且并发地检测离群区域,直到发现足够多离群点。实验结果表明,该算法能有效检测离群区域,并且准确给出局部离群性,克服了普通算法中易受坏邻居干扰和区域缺乏原子性的局限。

关键词: 空间离群点; 区域检测; Delaunay 三角网; 裁边
中图分类号: TP311.13;TP391 **文献标志码**: A **文章编号**: 1001-3695(2013)03-0783-04
doi:10.3969/j.issn.1001-3695.2013.03.036

Spatial outlier region detection algorithm based on cutting edge strategy

CHEN Tao, WAN Jia-qiang, ZHU Qing-sheng
(*College of Computer Science*, *Chongqing University*, *Chongqing* 400044, *China*)

Abstract: In order to detect the existed outlier regions in spatial data sets, this paper proposed a cutting-long-edges-based approach. It used Delaunay triangle mesh to construct adjacency relationships. Then it according to non-spatial attributes addressed the difference between nodes. Finaly it iteratively cut long edges, and concurrently detected anomalous regions, and it found till enough outlier. Experiments demonstrate that out approach can effectively discover anomalous regions and the abnormality measure can reflect real local difference. Especially, the approach overcomes the general drawbacks such as the disturbance of bad neighbors and the lack of atomicity of regions.

Key words: spatial outliers; region detection; Delaunay triangle; cutting edges

空间离群区域检测,即利用空间信息确定节点拓扑结构,并在此拓扑结构中发现异常区域。空间离群区域检测在空间数据库中有着重要的应用,主要包括气候分析、环境科学以及地球科学等领域[1]。对于传统的离群检测算法,如 COF[2]、LOF[3]等,主要是寻找远离主体的离群点。这些算法不关心数据分布的拓扑结构,因此不适合用于空间离群检测。空间离群检测采用二分算法的思想,典型的有 SLZ[4]、SLOM[5]、SLOF[6]。这些算法将空间数据的属性划分为空间属性与非空间属性。空间属性用来确定对象的邻接关系;非空间属性则用来度量对象的相似性或差异性,从而计算出每个空间对象的离群得分。这些算法主要针对空间离群点检测。但是,当空间离群点相互靠近而形成离群区域时,以上算法却性能低下。由于这些算法采用局部偏差的统计量来度量每个点的离群程度,容易受到离群区域中各个对象相互包庇的影响,因此区域中的离群点极可能被误判为正常数据。而在许多实际应用中,空间离群常常以区域的形式出现[7]。因此,空间离群检测除了要检测单点,更多地需要用于发现成片的离群区域。

离群区域是指与外部邻居的差异较大而内部对象较为相似的小的成片的空间对象集合。目前,对区域进行离群检测的研究较少。文献[8]提出一种静态扫描的算法,然而它并不能发现任意形状的离群区域。文献[9]提出一种异常窗口(区域)检测算法——FS3。该算法利用随机走(random walks)的思想来发现区域。它将窗口容量小于指定阈值的区域标记为离群区域。FS3 算法能够识别不规则形状的离群区域,但无法检测离群单点。文献[10]提出一种基于 KNN 图的空间离群区域检测算法 ROD,该算法通过删去图中所有权重小于阈值 T_{EDGE} 的边来寻找离群区域。ROD 能够同时检测单点与区域,但由于算法要求的参数较多,特别是最少邻居数 k 和阈值 T_{EDGE} 的选择比较困难,为了获得满意结果,ROD 需要反复调整参数,而每次对参数的调整都会导致 KNN 图的重建。因此,ROD 算法的参数依赖性高,检测结果也不稳定。区域检测的一个难点是区域范围的确定。离群区域检测必须确保所发现异常区域的原子性,即每个异常区域应该恰好覆盖一个成片的离群区域。当前的方法在这方面仍然存在很多不足。

1 相关定义

定义 1 空间近邻。给定空间数据集 X,图 $G=\langle V,E\rangle$ 为生成相应的 Delaunay 三角网格,其中 V 为顶点集,E 为边集。若空间对象 $x,y\in V$,边 $e\in E$,使得 x 与 y 间通过边 e 直接相连,则称 x 与 y 互为空间近邻,记为 $SN(x,e,y)=\text{true}$。需要指出的是,此处的近邻是指两个对象存在一条边连通,即存在邻接关系,而非单纯距离上的靠近。

定义 2 异质性边权。给定空间数据集 X,空间对象 $x\in X,y\in X$,若 $SN(x,e,y)=\text{true}$,则将 x 与 y 之间的异质性边权记为 W_e。

$$W_e=\frac{\|x-y\|}{(\|x\|+\|y\|)/2}$$

其中:$\|\cdot\|$ 为欧几里德范式。W_e 的大小反映了 $e=\overline{xy}$ 所关联的近邻对象 x 与 y 之间非空属性的差异程度,W_e 越大,说明近邻对象之间的差异越显著,表现出的空间异质性也越强。

收稿日期: 2012-07-23; **修回日期**: 2012-09-06 **基金项目**: 国家自然科学基金资助项目(61073058)

作者简介:陈涛(1988-),男,湖南耒阳人,硕士研究生,主要研究方向为数据挖掘(perfe_ct@163.com);万家强(1982-),男,重庆人,博士研究生,主要研究方向为数据挖掘;朱庆生(1956-),男,安徽当涂人,教授,博导,博士,主要研究方向为数据挖掘、图像处理、数字农业.

定义 3 WDNN 图。图 $G=\langle V,E\rangle$ 为给定空间数据集 X 所生成的 Delaunay 三角网格,若任意 $e\in E$ 被添加上异质性边权 W_e,则称图 G 为 WDNN 图。

定义 4 区域。给定一个连通图——WDNN 图 G,SG 为图 G 通过裁边策略得到的子图集合,对于任意连通子图或单点 $R\subset SG$,称 R 为区域。

定义 5 离群区域。给定区域 R 与区域阈值 T_R,$|R|$ 表示区域 R 的势,即该区域包含样本数量,若 $|R|<T_R$,称 R 为离群区域,记做 $\mathrm{outlier}(R)=\mathrm{true}$。

特别地,当 $T_R=1$ 时,表示该区域只是一个离群单点。

定义 6 邻接区域。给定 WDNN 图 $G=\langle V,E\rangle$,区域 R_1,$R_2\subset G$,若存在对象 $x\in R_1$,对象 $y\in R_2$ 与边 $e\in E$ 满足 $SN(x,e,y)=\mathrm{true}$,则称 R_1 与 R_2 互为邻接区域,记做 $\mathrm{Nbr}(R_1,R_2)=\mathrm{true}$。

定义 7 邻接偏差系数。给定 WDNN 图 G 和区域阈值 T_R,经过一定裁剪后得到 $G'=\langle V,E\rangle$,一个连通子图 $R\subset G'$,且 $|R|<T_R$,R 的邻接区域集为 $S_R=\{R'\mid R'\subset G',\mathrm{Nbr}(R,R')=\mathrm{true},|R'|=1\}$,$R$ 与 S_R 构成二部图,其边集为 $E'=\{e\mid e\in E,SN(x,e,y)=\mathrm{true},x\in R,y\in S_R\}$,则定义 R 的邻接偏差系数为

$$K_{\mathrm{SOF}}(R)=\min(E')$$

检测区域与其邻接区域之间构建一个二部图(此二部图的边权由 WDNN 图给出),二部图的最短边即为邻接偏差系数 $K_{\mathrm{SOF}}(R)$。$K_{\mathrm{SOF}}(R)$ 的值越大,说明 R 与周围正常区域之间的偏差越大,R 所表现出的离群程度也就越高。

2 基于裁边策略的空间离群检测算法

2.1 算法主要思想及步骤

本文采用最大边优先截断的裁边策略,能够尽快分离具有最大差异的空间近邻对象,而将相对稳定的区域保留下来,算法的任务就是在这些稳定的区域中找出感兴趣的离群区域。其思想为:反复裁去当前 WDNN 图中的最大边,且每次裁边均需检查是否产生小的子图,即离群区域;再将小的子图与其邻接区域构成二部图,此二部图的最短边恰好描述出该子图的异质性;接着移除小子图,进行下一次裁边,直到检测到的小子图总量包含 10% 的样本为止(笔者认为离群点属于少数,其总量一般少于 10%)。具体步骤如下:

输入:样本空间属性集 $S=\{s_1,s_2,\cdots,s_n\}$;非空间属性集 $X=\{x_1,x_2,\cdots,x_n\}$;区域阈值 T_R。

输出:离群区域集 O。

a)构建 WDNN 图。利用空间数据集 S 建立 Delaunay 三角网格,即建立邻接图;再用非空间属性集 X 为邻接图赋上边权。

b)排序 WDNN 图边。将 WDNN 图的边集队列 E 按边的权重大小进行降序排序。

c)裁边并寻找离群区域。在 WDNN 图中迭代地裁去边集队列中权重最大的边,且每次裁边均检索出对象数目小于 T_R 的区域,将其标记为离群区域,并归入 O 集;同时需要为找到的小区域打分,即计算其在当前图中的邻接偏差系数 $K_{\mathrm{SOF}}(R)$;接着从当前图中移除找到的小区域,并进入下一次迭代裁边,直到找到的离群集合 O 包含 10% 的样本。

d)输出离群区域。将 O 中离群区域 R 按 $K_{\mathrm{SOF}}(R)$ 降序排列,并输出结果。

2.2 WDNN 图的构建

WDNN 图是一个无向加权连通图,用于确定各个对象节点的关系。WNDD 图的构建包括两个步骤:

a)确定邻接关系。空间对象(节点)的邻接情况是空间离群检测的一个关键环节。通常可以采用 KNN 查询的方式来确定邻接点,即直接将空间属性最相似的 k 个邻居当做邻接对象。然而现实应用中,空间属性最相似的点不一定就是邻接点,例如网络布局、全国的省区分布等,这些节点的邻接关系并不以空间距离为准,而是以区域划分或者现实需要为准。Delaunay 三角网格恰好能描述这种划分布局下的邻接关系。Voronoi 图[10]是一种区域划分图,它能将 2 维平面的各个节点用凸多边形划分成不同的区域。而 Delaunay 三角网格是 Voronoi 图的对偶图,即是 Voronoi 图划分下的邻接图。由于 Delaunay 三角网格能很好地描述现实中的邻接关系,所以文中采用 Delaunay 三角网格来建立邻接图。关于 Delaunay 三角网格的构建,通常需要一个小角度来约束三角网格的形成,以建立更加合理的邻接关系[12]。

b)确定邻接点间差异性,即给出邻接节点间的偏差度量。对于邻接的一对节点,其差异性由定义 2 给出。这里的差异性采用比值而不是绝对值。通常情况下,绝对差值被用做偏差度量。然而绝对差值对于差异性描述缺乏可比性,如 5 和 10 的差异性明显强于 100 和 106,其绝对差值却会给出相反的结论。所以,本文设计了定义 2 中的异质性度量来表示邻接边权,且令不邻接边权为无穷大(inf),从而得到带权 Delaunay 图——WDNN 图。

2.3 裁边及区域检测

裁边过程是一个迭代的过程。本文算法每次只裁掉当前图中最长的边,并检查是否有离群区域,然后移除离群区域,接着进入下一次迭代裁边。每次迭代包括三个环节:

a)裁边,即反复裁掉当前图的最大边。在这个过程中,需要一边裁边一边检测离群区域,即每裁一边均需检测剩下的图中是否存在离群区域。ROD 也是基于裁边的检测方法。其做法是直接裁掉足够多的长边,然后检测离群区域。其缺陷是裁边的终止条件难以确定。因此,ROD 会因不同的终止条件得出不同的检测结果,从而导致 ROD 算法性能不稳定。笔者认为离群的区域总是由于大边的逐渐裁剪而逐渐被分离出来。所以本文采用的裁边策略是每次只裁去最大边并同时作一次离群区域检查,接着移除离群区域后进入下次裁边。

b)检测离群区域,其过程即为每次裁边后寻找小的连通子图。笔者认为子图 R 的势小于阈值 T_R 即为离群区域(定义 5)。这个环节就是根据裁边后的连通情况寻找小区域。一般情况下可以采用聚类的方法找出小区域。然而聚类的时间复杂度一般较高,多为 $O(N^2)$。所以,将连通图存入邻接表中,只需要一次遍历即可找到小于阈值 T_R 的区域,即每次裁边后的离群查找时间复杂度为 $O(N)$。

c)度量,即给离群区域打分。本文使用二部图短边度量区域离群性,考察离群区域与其他离群区域的连通情况。每次裁边最多产生两个小于阈值 T_R 的子图。如果某一次裁边只产生一个离群区域 R,那么根据定义 7 求解出邻接偏差系数 $K_{\mathrm{SOF}}(R)$,并作为离群分值 $DO(R)$。如果某一次裁边产生两个离群区域 R_1 和 R_2,则 $DO(R_1)=DO(R_2)=K_{\mathrm{SOF}}(R_1+R_2)$。第二种情况是一种临界状态,属于特殊情况。由于某条长边的裁剪,恰好某个大于阈值 T_R 的子图被划分为两个离群区域 R_1 和 R_2,那么说明 R_1 和 R_2 是连通的,且当前切割的长边即为当前两个离群区域的最短边。然而这条短边的权重无法反映出真实离群程度。所以,本文采用 $K_{\mathrm{SOF}}(R_1+R_2)$ 表示 R_1 和 R_2 在当前图中的离群程度。

2.4 算法描述

本文算法的伪代码描述如下:

输入:样本空间属性集 $S=\{s_1,s_2,\cdots,s_n\}$;非空间属性集 $X=\{x_1,$

```
x_2,…,x_n};区域阈值 T_R。
    输出:离群区域集 O;离群分值 DO。
    计算 Delaunay 三角网格:
      TRI_{n×3} = DelaunayTri(S);
      /TRI 的每一行代表一个三角形,所有三角形的边最终形成一个连通图 G = 〈V,E〉。
    计算 WDNN = [wr(x_i,x_j)]_{n×n};
      /其中 SN(x_i,e,x_j) = true,则 wr(x_i,x_j) = w_e;否则 wr(x_i,x_j) = inf
    降序排列 WDNN 的边集为 QUE;
    count = 0;设置两个当前图 CG = G 和 G' = G;
    while(count < n * 0.1)
    {
      QUE 出栈当前最长边 pq;
      修改当前图 CG,使得 SN(p,e,q) = false;
    遍历 CG 得到所有连通子图,得到如下:
      G' = {g_1,g_2,…,g_l}, |g_1| <= |g_2|…|g_l|;
      if |g_1| <= T_R && |g_2| > T_R
      {count = count + |g_1|;
        将 g_1 保存到 O 中;
        DO(g_1) = K_SOF(g_1);
        /K_SOF(g_1)为 G' 中偏差值
        从 CG 和 G' 中均移除 g_1;
    } else
    {if |g_1| <= T_R && |g_2| <= T_R
        {将 g_1、g_2 保存到 O 中;
        if K_SOF(g_1 + g_2) == inf
          {DO(g_1) = DO(g_2) = K_SOF(g_1 + g_2);}
        /K_SOF(g_1 + g_2)为 G 中偏差系数
        else
          {DO(g_1) = DO(g_2) = K_SOF(g_1 + g_2);}
        /K_SOF(g_1 + g_2)为 G' 中偏差系数
        从 CG 和 G' 中均移除 g_1、g_2;
    }}
        /裁剪后的 CG 和 G' 进入下次裁边
    }/end-while
    [DO,tidx] = sort(DO,2,'descend');
    /对离群区域按分值降序排列
    O = O(tidx);
```

2.5　时间复杂度分析

设数据集规模为 N,本文算法时间复杂度如下:

步骤 a)中,采用分治算法生成 Delaunay 三角网的最坏时间复杂度为 $O(N\log N)$,构造边权的时间复杂度为 $O(|E|)$。由于 WDNN 图的边集最大数量为 $3|V|-6$,$O(|E|)=O(3N-6)$,该项总的时间复杂度为 $O(N\log N+3N-6)\approx O(N\log N)$。

步骤 b)中,采用堆排序的时间复杂度可以达到 $O(|E|\log|E|)=O((3N-6)\log(3N-6))\approx O(N\log N)$。

步骤 c)中,为寻找离群区域,遍历裁剪后的 WDNN 图,其时间复杂度为 $O(N)$;另外 Delaunay 三角网格中 $|E|\times 0.1<N$,可认为其迭代裁边时间复杂度近似为 N,则步骤 c)时间复杂度为 $O(N^2)$。

步骤 d)中,排序的时间复杂度为 $O((0.1\times N)\log(0.1\times N))$。

根据上述分析,整个算法时间复杂度为 $O(N^2)$。

3　实验及分析

关于离群区域检测,本文的主要工作是区域划分和离群度量。通过与 FS3[9]、ROD[10]的对比实验,将对比分析本文算法在离群区域划分和离群度量的性能。相关实验数据包括一组人工数据集和三组真实数据集。图 1 给出了人工数据集;而真实数据集 1、2、3 分别来自国内外国家统计局。数据集 1 来自美国统计局,是关于美国各个州性别比例的数据,包括 51 个样本;数据集 2 是关于中国大中城市房价走势的数据,包括 70 个样本,来自中国统计局数据发布网站;数据集 3 为中国统计局发布的人口统计数据,包括全国所有区县的人口统计信息。实验中,各地区的经纬度作为空间属性;其他信息分别源自以上数据集,且作为非空间属性。另外,由于个别区域地理位置远离主体,不便于图示,所以在实验中被移除。其中,数据集 1 中的 Alaska 和 Hawaii 被移除,数据集 2 中的乌鲁木齐被移除。

3.1　人工数据集实验

为了检验本文提出算法的可行性,将首先进行人工数据集实验,然后在 3.2 节中利用真实数据集验证文中算法的优越性。图 1 所示的人工数据集中,各个数值即非空间属性,其坐标即空间分布信息。本文算法结果为:$R_1=[a(1,5),a(1,4),a(1,3),a(2,3),a(2,4),a(2,5)]$,通过观察可知该区域是一个离群区域;$R_2=[a(6,1)]$为离群单点。FS3 算法能够容易地找到 R_1。然而由于 FS3 不擅长寻找离群单点,它将 $a(6,1)$融合到正常区域,从而丢失了离群区域 R_2。关于 ROD 算法,其检测思想主要是寻找小的连通子图,因此它能顺利找到 R_2。然而 ROD 的裁边阈值难以确定,根据不同的阈值可能得出不同的离群区域,即 R_1 可能被分解,从而破坏了离群区域的原子性。加上 ROD 的离群度量是基于统计的偏差,因此 ROD 很容易因为错误度量丢失一些离群区域。例如 $a(1,5)$和 $a(1,4)$与其他邻居离群点差异较小,即 $R_3=[a(1,5),a(1,4)]$相对于邻居点 $a(1,3)$,$a(2,3)$,$a(2,4)$,$a(2,5)$的差异性较低。当 ROD 阈值过小时(为了检测出 $a(6,1)$,必须选择小阈值),R_3 由于低离群性而被丢失。

y \ a(x,y)	1	2	3	4	5	6
5	1	1.2	3	3	3	3
4	1	1.2	3	3	3	3
3	1.1	1.1	3	3	3	3
2	3	3	3	3	3	3
1	3	3	3	3	3	2

x

图1　人工数据集

从人工数据集实验可以看出,FS3 不善于单点检测;ROD 容易破坏区域的原子性,导致不稳定的检测结果。

3.2　真实数据集实验

针对三组真实数据集,本文算法以及 FS3、ROD 的实验结果如表 1 ~ 3 所示,其中各区域均按离群分值降序排列,[ ]表示一个连通成片的区域。

表 1　本文算法检测结果

数据集	离群区域
1	{DC,[CO,ID,MT,UT]}
2	{温州,[兰州,西宁],南充,泉州}
3	{西沙群岛,孟连傣族拉祜族佤族自治县,石龙区,北站区,泸水县,[肃北蒙古族自治县,阿克塞哈萨克族自治县]}

表 2　FS3 检测结果

数据集	离群区域
1	{[ID, MT, ND, WY], [MN,WI]}
2	——
3	{[孟连傣族拉祜族佤族自治县,镇康县,耿马傣族佤族自治县,沧源佤族自治县]}

表 3　ROD 检测结果

数据集	离群区域
1	{DC, *}
2	{温州,[赣州,岳阳],[长沙,常德]}
3	{西沙群岛,孟连傣族拉祜族佤族自治县,石龙区,北站区,泸水县,[肃北蒙古族自治县,阿克塞哈萨克族自治县]}

图 2 和 3 分别描述了本文算法检测出数据集 1、2 的相应离群区域。另外,由于数据集 3 样本过大,无法图示,因此略去其图示检测结果。

在数据集 1 中,DC 地区的男女比例是 0.895,明显低于周边地区(平均比率为 0.97),因此该区域为一个明显的单点离群区域。然而 FS3 将[DC,DE,MD]当做一个区域,并且划分

为正常区域,通过观察可发现,DC(0.895)和 DE(0.939)、MD(0.936)的相似程度并不高,因此这个区域划分并不合理。另外,FS3 也未能发现[CO,ID,MT,UT],而是给出最离群区域[ID,MT,ND,WY]。经过查看各个地区男女比例,发现 ID(1.004)、MT(1.008)、ND(1.021)、WY(1.041)并不完全相似,特别是[ND,WY]与[ID,MT]相差较大,由此可见 FS3 的区域划分缺乏合理性。而 ROD 除了成功发现离群区域 DC 外,要确定其他离群区域较为困难。由于涉及参数过多(4 个),并且轻微的参数变化导致不同的检验结果,特别是未能给出一个较好的终止条件,最终导致 ROD 算法结论的不确定性。从实际检测结果可知,本文算法所检测出的离群区域[CO(1.005),ID(1.004),MT(1.008),UT(1.009)]中各个州的性别比例高度相似,更符合区域划分原则。

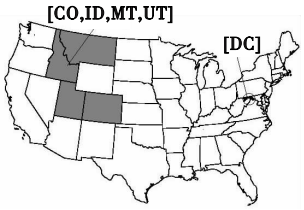

图2 本文算法在数据集1中检测出的离群区域分布

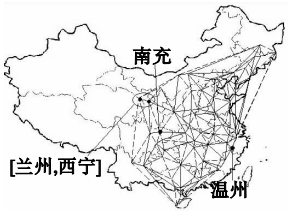

图3 本文算法在数据集2中检测出的离群区域分布

数据集 2 中,由于 FS3 并不适用于高维数据,因此不予考虑。ROD 的检测结果是{温州,[赣州,岳阳],[长沙,常德]}。根据离群区域内各个子区域应高度相似的原则,可知长沙应与常德相似而与赣州相异。实际上,长沙与常德的非空间属性差值(欧式距离)为 3.55,而与赣州的非空间属性差值为 1.28,显然 ROD 的区域划分不合理。

数据集 3 中,西沙群岛的性别比率为 3.7,孟连傣族拉祜族佤族自治县是为 1.69。由于 FS3 并不擅长单点检测,因此它无法检测西沙群岛。FS3 能检测到一个最离群区域为[孟连傣族拉祜族佤族自治县(1.69),镇康县(1.21),耿马傣族佤族自治县(1.25),沧源佤族自治县(1.30)]。但孟连傣族拉祜族佤族自治县是一个明显的单独离群区域,其与其他几个县存在明显的差异,然而 FS3 将该区域融合到其他几个县,造成区域冗余,从而导致错误的区域检测。由于数据集 3 为大数据集,且单个离群点较多,所以 ROD 算法与本文算法所检测到的前几个离群区域相同。

3.3 算法及实验分析

目前,空间离群检测的研究主要集中点检测方面,而关于离群区域检测的研究较少。FS3 和 ROD 算法为目前的区域检测算法。前者不适合单点检测,而后者容易破坏区域的原子性。FS3 采用随机走策略融合区域,单个离群点很容易被融合到非离群区域从而造成单点丢失;且随机走策略并不是最相似邻居融合的策略。因此,FS3 区域划分并不准确,会造成区域冗余或错分的现象。而 ROD 算法采用 KNN(k nearest neighbors)图表示邻接关系,并采用一次性裁边的方式得到离群区域。由于 KNN 图表示近邻关系而非邻接关系,所以很可能造成错误区域划分。此外,ROD 过渡依赖裁边阈值,不同的阈值往往得到不同的检测结果,从而导致检测结果的不稳定,也就无法确保离群区域的原子性。KNN 图能反映物理空间近邻关系,而不是邻接关系,一般使用者通常把近邻关系等价为邻接关系,从而导致错误的区域划分。现实中往往存在邻接但并不近邻的情况(如主干网络的拓扑结构),则导致 ROD 无法正确划分区域。再者 ROD 算法的离群度量采用基于局部偏差的统计量,离群小区域很可能相互包庇,从而导致离群区域的丢失。

根据实验可知,本文提出的空间离群区域检测算法在区域划分方面更为准确。该方法包含聚类的思想,能确保划分出的区域是内部高度相似的。为了描述邻接关系,该方法采用了 Delaunay 三角网格。Delaunay 三角网格作为 Voronoi 图的对偶图,是较为理想的邻接关系图。Voronoi 图是一种用于划分区域的方法,而 Delaunay 图正是描述其区域划分情况下的邻接图。现实中的离群区域划分主要考虑邻接关系,而 Delaunay 图正好切合此区域划分的要素。算法中的裁边策略是每裁一条长边均需检查是否有离群区域出现,而不是裁掉足够多的边以后再检查离群区域。由于离群区域总是随着大边的逐步裁剪依次出现,且算法不依赖裁边阈值,因此检测结果会相对稳定。即使参数的轻微变化,检验结果也不会发生巨大变化。另外,关于离群区域的度量,常规方法多采用统计量,因此无法避免坏邻居的错误投票,比如离群点相互为参考评价对象,则会相互包庇或者恶意投票。而本文方法采用二部图短边来度量区域离群性,其本质是考虑了连通性——离群区域与主体部分连通性差则表明为离群区域,反之为正常区域。综上几个特点,本文方法在区域划分和离群度量方面均有明显改进,且算法参数较少,实现简单,实用性强。

4 结束语

针对空间离群区域挖掘,本文提出一种新颖的空间离群检测算法。实验结果表明,该算法能有效地识别离群单点和离群区域;与 FS3、ROD 算法相比,算法在区域划分和离群度量方面均具有明显的优势。虽然本文算法与 ROD 算法同属于裁边策略,但文中的方法不依赖裁边阈值。文中提出的区域划分策略按照聚类的思想,将相似的离群点连通并移出,从而找到正确的离群区域。另外,本文采用二部图短边来度量区域的离群性,短边恰好表示了数据连通性,它有效地避免了离群点相互包庇的情况,使得离群度量更真实地反映出离群区域的偏离情况。

参考文献:

[1] SHEKHAR S, EVANS M R, KANG J M, *et al*. Identifying patterns in spatial information: a survey of methods[J]. Data Mining and Knowledge Discovery, 2011, 1(3): 193-214.

[2] TANG Jian, CHEN Zhi-xing, FU A W. Enhancing effectiveness of outlier detections for low density patterns[C]//Proc of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. London: Springer-Verlag, 2002: 535-548.

[3] BREUNIG M, KRIRGEL H, NG R T, *et al*. LOF: identifying density-based local outliers[C]//Proc of ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2000: 93-104.

[4] SHEKHAR S, LU C T, ZHANG Pu-sheng. A unified approach to spatial outliers detection[J]. Geoinformatica, 2003, 7(2): 139-166.

[5] SANJAY C, SUN Pei. SLOM: a new measure for local spatial outliers[J]. Knowledge and Information Systems, 2006, 9(4): 412-429.

[6] 薛安荣,鞠时光,何伟华,等. 局部离群点挖掘算法研究[J]. 计算机学报, 2007, 30(8): 1455-1463.

[7] LU C T, KOU Yu-feng, ZHAO Jiang, *et al*. Detecting and tracking regional outliers in meteorological data[J]. Information Sciences, 2007, 177(7): 1609-1632.

[8] NEILL D B, MOORE A W, PEREIRA F, *et al*. Detecting significant multidimensional spatial clusters[C]//Advances in Neural Information Processing Systems. 2005: 969-976.

[9] JANEJA V, ATLURI V. Random walks to identify anomalous free-form spatial scan windows[J]. Trans on Kownledge and Data Engineering, 2008, 20(10): 1378-1392.

[10] KOU Yu-feng, LU C T, RAIMUNDO F, *et al*. Spatial outlier detection: a graph-based approach[C]//Proc of the 19th IEEE International Conference on Tools with Artificial Intelligence. Paris: IEEE Press, 2007: 281-288.

[11] AURENHAMMER F. Voronoi diagrams: a survey of a fundamental geometric data structure[J]. ACM Computing Surveys, 1991, 23(3): 345-405.

[12] SHEWCHUK R. Delaunay refinement algorithms for triangular mesh generation[J]. Computational Geometry: Theory and Applications, 2002, 22(1): 21-74.