基于空间约束的离群点挖掘*)

薛安荣　鞠时光
(江苏大学计算机科学与通信工程学院　镇江 212013)

摘　要　由于现有的空间离群点检测算法没有很好地解决空间数据的自相关性和异质性约束问题，提出用计算邻域距离的方法解决空间自相关性约束问题，用计算空间局部离群系数的方法解决空间异质性约束问题。用离群系数表示对象的离群程度，并将离群系数按降序排列，取离群系数最大的前 m 个对象为离群点，据此提出基于空间约束的离群点挖掘算法。实验结果表明，所提算法比已有算法具有更高的检测精度、更低的用户依赖性和更高的效率。

关键词　空间局部离群系数，邻域距离，空间离群点，离群点检测

Outlier Mining Based on Spatial Constraint

XUE An-Rong　JU Shi-Guang
(School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang 212013)

Abstract　Major drawbacks of existing spatial outlier detection algorithms are that the spatial autocorrelation and spatial heterogeneity of spatial objects aren't considered, normal objects tend to be falsely detected as spatial outliers or true spatial outliers tend to be ignored. We define neighborhood distance to overcome spatial autocorrelation constraint and defined spatial local outlier factor (SLOF) to overcome spatial heterogeneity constraint. SLOF captures the local behavior of datum in their spatial neighborhood. SLOF-based algorithm of spatial outlier detection can successfully find local spatial outliers which appear to be meaningful, but can otherwise not be identified with existing approaches. The experimental results show that our algorithm outperforms other existing algorithms in detection accuracy, user dependency and efficiency.

Keywords　Spatial local outlier factor, Neighborhood distance, Spatial outlier

1　引言

离群点(outlier)检测是数据挖掘的基本任务之一[1,2]，其目的是消除噪音或发现潜在的、有意义的知识。到目前为止，还没有一个广为接受的离群点的正式定义，但 Hawkins 的定义抓住了概念的精髓："一个离群点是一个观察点，它偏离其它观察点如此之大以至引起怀疑是由不同机制生成的"[2]。典型的应用包括信用卡欺诈的检测和在电子商务中犯罪行为的监测。

离群点检测算法很多，可分为：基于分布的、基于深度的、基于距离的、基于密度的和基于聚类的。现在对离群点的研究主要集中在高维大数据量和基于条件约束的离群点检测，以及对离群原因的分析解释上。

基于空间约束的离群点检测算法的研究已经引起人们的兴趣，出现了一些算法。首先，在空间统计学中，出现了图形检测和代数检测两类方法，如变差云图(variogram cloud)法和 Z-Score 法。但这些方法由于没有考虑空间数据的特点，没有区分空间和非空间属性，其检测效果不佳。Shekhar、Lu 和 Zhang 学者首先提出将空间属性与非空间属性区分开来的二分算法[3]，并通过对象与其邻域的非空间属性值之差或之比，来消除空间的自相关性，并用该值表示对象与其邻域的偏差，我们简称该算法为 SLZ 算法。SLZ 算法未能很好地解决空间的异质性问题，因此，主要检测的还是全局离群点。Chawla 和 Sun 学者同时考虑了空间的自相关性和异质性，用欧拉距离来消除空间对象与其邻域间的自相关性，引入波动参数 β，并用 β 和对象与其邻域的欧拉距离的乘积表示空间局部离群度 SLOM(Spatial Local Outlier Measure)[4]。但由于 β 仅由对称分布状况来决定，在空间邻居较少或波动幅度较小的情况下难以准确表现波动情况，因此出现漏检和误检现象。为了解决空间异质性约束条件，我们用对象的欧拉距离与其邻域的平均欧拉距离之比，表示对象的空间局部离群程度，即空间局部离群系数 SLOF (Spatial Local Outlier Factor)。应用中，为了避免分母为 0 的情况，采取分子分母同时加上一个较小的正常数 δ，通过对 SLOF 排序，可挖掘出离群点。实验表明，基于 SLOF 的算法具有较高的准确检测率、较低的用户依赖性和较高的运算效率。

2　基于空间约束的离群点概念及 SLOF 计算

2.1　基于空间约束的离群点

空间对象的属性按性质可分为两类：空间属性和非空间属性。空间属性包括位置、形状、方向、空间邻接关系和其它几何或拓扑性质，非空间属性包括名称、年代、长度和高度等[5]。非空间维属性是对象固有的，从本质上刻画了数据对象，而空间维属性并非对象所固有，但提供了对象的位置索引。由于空间数据具有空间自相关性和空间的异质性[5,6]，因此，空间数据具有局部性特点，而空间邻居在空间数据的分析中扮演着重要角色。所谓空间自相关性(Spatial autocorrelation)是指每个空间对象的属性受它的空间邻居的影响，而

*)基金项目：国家自然科学基金(60373069)，江苏省高校自然科学基金(05KJB520017)。薛安荣　博士生，副教授；鞠时光　CCF 会员，博士，教授，博士生导师。

空间异质性(Spatial heterogeneity)是指不同地区的属性数据的变化趋势是不同的。

仅根据空间属性确定的离群点,如孤岛、偏僻的山村等,可应用已有的离群点检测算法完成,对此本文不作研究。本文将研究以空间关系为约束条件,以非空间属性为比较属性的离群点检测算法。已有的文献将基于空间约束的离群点称为空间离群点,在不引起混淆的情况下,我们也沿用这个概念。

空间离群点是指对象的非空间属性值与其空间邻居相比明显不同或异常。空间离群点检测在地理信息系统和空间数据库的很多应用中非常有用,这些应用包括公共安全、公共健康、生态环境、交通运输和基于位置的服务等领域[5]。

2.2 相关定义

假设对象集 $O=\{o_1,o_2,\cdots,o_n\}$,由 n 个对象组成,对象 $o\in O$ 的空间属性函数是 $s(o)$,非空间属性函数是 $f(o)$,$f(o)$ 的维度为 d-维,σ_c 表示在指定条件 c 下的空间邻接关系。d-维非空间属性 $f(o)$ 表示为 $(f(o_1),f(o_2),\cdots,f(o_d))$。

定义 1(空间邻居) 对象 o 的空间邻居是指与对象 o 在指定条件 c 下,存在空间邻接关系 σ_c 的对象。即 $\forall o\in O,\exists p\in O\backslash\{o\}$,使得 $s(p)\sigma_c s(o)$ 为真,则对象 p 是对象 o 的空间邻居。

定义 2(空间邻域) 对象 o 的空间邻域 $N(o)$ 是指对象 o 的所有空间邻居的集合,即 $\forall o\in O,N(o)=\{p\mid s(p)\sigma_c s(o)=\text{true},p\in O\backslash\{o\}\}$。

定义 3(加权距离) 设 $o_i,o_j\in O$,o_i 和 o_j 的 d-维非空间属性是 $f(o_i)$ 和 $f(o_j)$,其中 $f(o_{i_k})$ 和 $f(o_{j_k})$ 是第 $k(k=1,\cdots,d)$ 维规则化属性,且 $0\leqslant f(o_{i_k}),f(o_{j_k})\leqslant 1$,$w_k$ 是第 k 维的权值,且 $0\leqslant w_k\leqslant 1$,则数据对象 o_i 和 o_j 之间的加权距离为:

$$dist(o_i,o_j,w)=\sqrt{\sum_{k=1}^{d}w_k(f(o_{i_k})-f(o_{j_k}))^2},\ \sum_{k=1}^{d}w_k=1 \tag{1}$$

值得注意的是,这里的对象间距离不是对象间的空间距离,而是对象间的 d-维非空间属性距离。根据分析需要,如果不同属性对分析目标的贡献程度不同,则分配的权值也不同,贡献率大的权值大,反之则小,权值一般由领域专家决定。对象间的加权距离的计算:一方面消除了对象间的相关性;另一方面决定两个对象间的偏差,距离越大,对象间的偏差越大。

定义 4(邻域距离) 邻域距离是指对象 o 与空间邻域中所有对象的加权距离的平均值,即:

$$dist(o,N(o),w)=\frac{\sum_{p\in N(o)}dist(p,o,w)}{|N(o)|} \tag{2}$$

邻域距离表示对象与其邻域在非空间属性上的偏差,邻域距离越大,偏差越大,其离群程度越高。如果将所有对象的邻域距离按降序排列,则邻域距离最高的 m 个对象就是所要检测的 m 个离群点,这是全局意义上的离群点。

由离群点定义可知,对象与邻域中离群点的距离最大,为了消除邻域中离群点对邻域距离计算的影响,避免因离群点的影响致使正常数据被误检为离群点,剔除邻域中与对象的最大距离,因此修改(2)式为:

$$dist(o,N(o),w)=\frac{\sum_{p\in N(o)}dist(p,o,w)-\max\{dist(p,o,w)\mid p\in N(o)\}}{|N(o)|-1} \tag{3}$$

邻域距离代表了对象与其邻域的偏差,将邻域距离与其空间邻居进行比较得到对象在局部空间上的偏离程度,即空间局部离群系数。

定义 5(空间局部离群系数)对象 o 的空间局部离群系数定义为:

$$SLOF(o)=\frac{dist(o,N(o),w)}{\frac{\sum_{p\in N(o)}dist(p,N(p),w)}{|N(o)|}} \tag{4}$$

为了避免 SLOF 计算中分母为 0 的情况,设 δ 为非常小的正数,分子、分母同时加上 δ,则(4)式修改为:

$$SLOF(o)=\frac{dist(o,N(o),w)+\delta}{\frac{\sum_{p\in N(o)}dist(p,N(p),w)}{|N(o)|}+\delta} \tag{5}$$

SLOF 表示对象在局部空间上的离群程度,计算所有对象的 SLOF,并按降序排列,离群度最大的前 m 个对象就是所求的空间离群点。可以证明只要 δ 取足够小,就能保证加 δ 后不改变 SLOF 的原有顺序,限于篇幅这里省去证明。

这样利用邻域距离解决了空间自相关性问题,利用 SLOF 解决了空间异质性问题,而利用 SLOF 的顺序解决了离群点的判断问题。

由于(1)~(4)式的计算中,所有非空间属性均规则化到[0,1]区间上,因此有 $\frac{\delta}{d+\delta}\leqslant SLOF(o)\leqslant\frac{d+\delta}{\delta}$,$\delta$ 的取值将决定 SLOF 的取值范围,但 δ 只要足够小,就不影响 SLOF 的顺序。当 $SLOF(o)\leqslant 1$ 时,对象 o 是正常对象,随着 SLOF 值的增大,其离群度增大,只有当 $SLOF>1$ 时,对象才可能是离群点。

3 基于空间约束的离群点检测算法

3.1 问题描述

给定:一个具有多维非空间属性的大的空间数据库,即 n 个空间对象 $O=\{o_1,o_2,\cdots,o_n\}$,空间邻居关系 $N=(N(o_1),N(o_2),\cdots,N(o_n))$,$N\subseteq O\times O$,对象 o_i 的空间邻域为 $N(o_i)(i=1,2,\cdots,n)$,空间对象 o_i,o_j,如果满足 $(o_i,o_j)\in N,i\neq j$,则 o_i,o_j 是空间邻居;对象 $o_i(s(o_i),f(o_i))$ 的空间属性是 $s(o_i)$,d-维非空间属性 $f(o_i)$ 表示为 $(f(o_{i1}),f(o_{i2}),\cdots,f(o_{id}))$。

目标:设计一个离群度的计算模型,根据该模型计算每个对象的离群度并按降序排列,输出离群度最大的前 m 个对象。

约束:每个空间对象具有空间属性和非空间属性,非空间属性是对象固有的,但非空间属性受其空间位置影响,即受空间自相关性和空间异质性的约束。

3.2 算法描述

Input:对象集 $O=\{o_1,o_2,\cdots,o_n\}$;对象 $o_i(s(o_i),f(o_i))$ 的 d-维非空间属性 $f(o_i)$ 表示为 $(f(o_{i1}),f(o_{i2}),\cdots,f(o_{id}))$;$\sigma_c$ 表示在指定条件 c 下的空间邻接关系;w_k 是第 k 维的权值

Output:离群点集。

Algorithm Based on SLOF:

```
(1)  for(i=1;i≤|O| ;i++){//|O|为O中的对象数量
(2)    o_i=Get_One_Object(i,O);//从O中选取一个对象
(3)    N(o_i)=Find_Neighbor_Nodes_Set(o_i,O,σ_c);//从O中寻
       找与o_i存在空间关系σ_c的所有邻居
(4)  }
(5)  for(j=1;j≤d;j++){
(6)    max_j=max(f(o_1j),f(o_2j),…,f(o_nj));//求第j维属性的最
       大值
(7)    min_j=min(f(o_1j),f(o_2j),…,f(o_nj));//求第j维属性的最
       小值
(8)  }
(9)  for(i=1;i≤|O| ;i++)
```

```
(10)    for(j=1;j≤d;j ++)
(11)      f(o_ij)=(f(o_ij)-min_j)/(max_j-min_j);
(12) for(i=1;i≤|O|; i ++){//计算对象与其邻域距离
(13)    for(dist(o_i)=0,maxdist=0,k=1;k≤|N(o_i)|;k ++){
(14)      for(tempdist=0,j=1;j≤d;j ++)
(15)        tempdist+=w_i^* (f(o_ij)-f(o_kj))^2;
(16)      tempdist+= sqrt(tempdist);
(17)      maxdist=max(maxdist,tempdist);
(18)      dist(o_i)+= tempdist;
(19)    }
(20)    dist(o_i)=(dist(o_i)-maxdist)/(|N(o_i)|-1);
(21) }
(22) for(i=1;i≤|O|;i ++){//计算对象的空间局部离群系数
(23)    for(ndist(o_i)=0,k=1;k≤|N(o_i)|;k ++)
(24)      ndist(o_i)+= dist(o_k);
(25)    slof(o_i)=(dist(o_i)+δ)/(ndist(o_i)/|N(o_i)|+δ);
(26) }
(27) Sort(slof);//对 slof 按降序排列
(28) Top_m_Set=Get_Top_m(slof,O);//取前 m 个对象
(29) Output(Top_m_Set);//输出离群对象
     End
```

3.3 算法分析

在上述算法中，邻域的确定是非常费时的，但由于是空间数据，利用空间特性及空间索引 R*-树来确定空间邻域，其计算复杂度大为降低。假设空间数据对象数目为 n，非空间属性维度为 d 维，对象的邻居数为 k(k 可变)，s 为 R*-树中每个索引结点的最少项数，则：语句(1)～(4)确定空间邻居，计算复杂度为 $O(n(k\log_s n))$；语句(5)～(11)规则化非空间属性为[0,1]，其复杂度为 $O(dn)$；语句(12)～(21)计算对象与其邻域距离，其复杂度为 $O(n(k\log_s n+kd))$；语句(22)～(26)计算 SLOF 的复杂度是 $O(n(k\log_s n))$；语句(27)排序的计算复杂度为 $O(n\log_2 n)$；语句(28)～(29)取前 m 个离群对象并输出的复杂度是 $O(m)$。故总的复杂度为 $O(n(k\log_s n+kd+\log_2 n))$。

4 实验测试及分析

我们用 VC++ 编写了以 R*-tree 为索引结构的 SLOF 算法程序，运行机器配置为 P4 1.66G CPU、512M 内存、操作系统为 Windows 2000、数据库为 SQL Sever 2000，测试采用文[4]中的合成数据和美国 2000 年的人口统计数据，限于篇幅，这里仅给出合成数据的测试结果。合成数据如表 1[4]，表 1 中有 100 个数据点，直接相邻的为邻居，所有邻居构成邻域。测试是从 100 个数据对象中检测出 4 个离群点。表 2 给出了我们所提出的 SLOF 方法($\delta=0.01$)以及引言中提到的 SLOM 方法和 SLZ 方法的测试结果。图 1 给出了三种方法检测出的离群点及其邻域数据波动状况，从图 1 可以看出，图 1 的(a)～(d)中离群点邻域波动比较平稳，而(e)～(f)的邻域波动较大，因此从图 1 可以看出(a)～(d)是最离群的 4 个离群点，其中(a)的邻域最为稳定，因此是最离群的。从表 3 可以看出，SLOF 算法检测的正确明显优于其它两种算法，运行时间与 SLZ 接近。虽然 SLOM 方法和 SLZ 方法均检测出 4 个离群点中的 3 个，但 SLOM 方法检测出(a)中的离群点，而 SLZ 方法未检测出，如果根据图示分析分别给图 1 中(a)～(d)的离群点分配权值 0.4～0.1，那么 SLOF 的检测的正确率为 100%，SLOM 为 70%，SLZ 为 60%，所以从检测正确率上来看 SLOF>SLOM>SLZ；从运行时间上来看，SLZ 与 SLOF 用时接近，SLOM 用时最多，如果采用在 SLZ 算法中采用平均值或中心值代替邻域中最高值计算重新计算，那么 SLZ 耗时最多，所以 SLOF>SLOM>SLZ。从对用户依赖性角度来看，SLZ 算法需要用户指定阈值，而 SLOF 和 SLOM 不需要，所以 SLOF 与 SLOM 优于 SLZ。

表 1 合成数据比较

	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	8	0	0	5	6	1	1	−9	3
8	0	0	0	0	−9	−8	−1	−2	9	5
9	0	0	0	0	19	−1	−2	−7	−3	−12

表 2 测试结果比较

序号	SLOF 方法		SLOM 方法		SLZ 方法	
	位置	值	位置	值	位置	值
1	(8,1)	18.4	(0,5)	0.428	(0,7)	24.0
2	(9,4)	3.88	(2,5)	0.248	(0,5)	23.6
3	(0,5)	3.03	(0,7)	0.206	(6,4)	−23.0
4	(0,7)	2.69	(8,1)	0.174	(9,4)	22.6

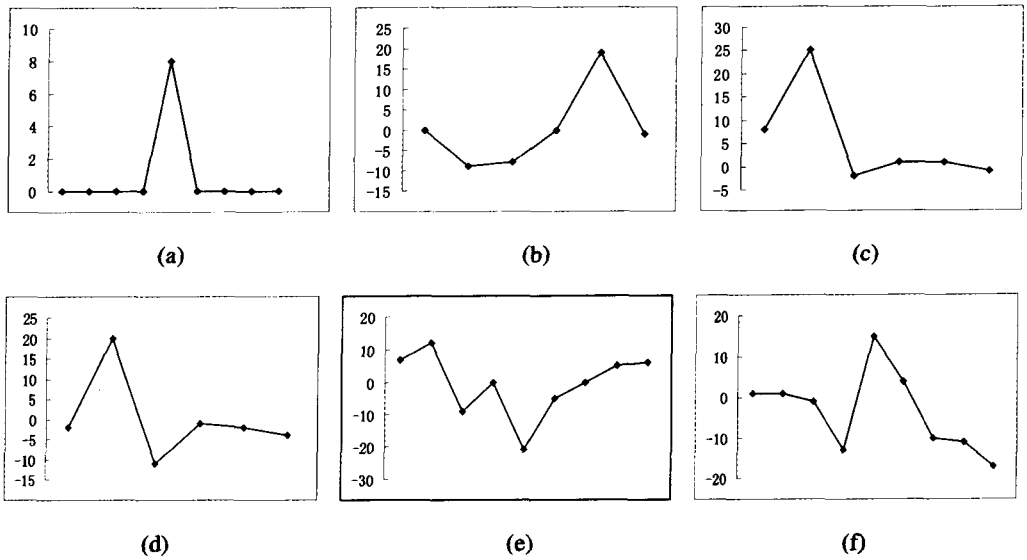

图 1 数据点与其邻居的波动图

(下转第 230 页)

为了比较，本文利用了两个不同的方法来处理同一个数据集。第一个简称为CRIA算法，包含了文[12]的属性约简算法和文[13]的属性值约简算法，其试验结果见表3；第二个是采用误差反传前馈神经网络(BPNN)，其网络模型为180-4-1，即输入层、隐含层和输出层节点数分别为180、4和1，学习算法采用反向传播算法，激活函数为Sigmoid函数，学习率和动量因子分别为0.2和0.9。试验结果见表4。

从列举在表2、3和4中的检测结果可看出，表2中的平均误判率(约24.31%)很接近于表4中的平均误判率(约22.43%)，并且远远低于表3中的平均误判率(约33.52%)。这说明了，与CRIA算法相比，由本文提出的基于粗集的T细胞表位预测方法得到的规则集具有较强的泛化能力。与BPNN算法的"黑箱性"相比，本文方法能在保证预测性能的前提下，提取出易于专家理解的产生式规则。

表4 BPNN算法的检测结果

实验序号	非结合肽(%)	低亲和力(%)	中亲和力(%)	高亲和力(%)	平均(%)
1	11.03	63.27	73.91	44.83	23.30
2	10.67	77.55	78.26	37.93	23.17
3	9.77	83.67	78.26	36.21	22.64
4	9.77	85.71	80.44	34.48	22.64
5	9.77	69.39	78.26	41.38	22.51
6	9.95	81.63	73.91	32.76	21.86
7	9.22	77.55	73.91	37.07	21.73
8	8.50	79.59	80.44	36.21	21.60
平均(%)	9.83	77.30	77.17	37.61	22.43

结论 为了解决现有的基于神经网络的T细胞表位预测模型所固有的"黑箱性"问题，本文巧妙地将T细胞表位预测领域知识融入到基于粗集理论的知识获取方法中，提出了基于粗集的T细胞表位预测方法。实验结果表明：本文提出的方法具有比CRIA方法更强的泛化能力；而与基于神经网络的T细胞表位预测方法相比，本文方法能在大致不降低预测精度的前提下，获取易于专家理解的产生式规则。这些规则有助于生物学专家将其注意力集中于某些很可能的关键模式上，并便于生物学专家通过对这些很可能的关键模式的验证和分析来进一步理解蕴含于其中的免疫学机理。

参考文献

1 陈慰峰. 医学免疫学. 北京：第三版. 人民卫生出版社，2000

2 Markus S, Toni W, Stefan S. Combining computer algorithms with experimental approaches permits the rapid and accurate identification of T cell epitopes from defined antigens[J]. Journal of Immunological Methods, 2001, 257: 1～16

3 Gulukota K, Sidney J, Sette A, et al. Two complementary methods for predicting peptides binding major histocompatibility complex molecules. Journal of Molecular Biology, 1997, 26: 1258～1267

4 Brusic V, George R, Margo H, et al. Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network[J]. Bioinformatics, 1998, 14: 121～130

5 Kun Y, Petrovsky N, Schonbach C, et al. Methods for prediction of peptide binding to MHC molecules: a comparative study. Mol Med, 2002, 8(3): 137～48

6 Pierre D, Arne E. Prediction of MHC class I binding peptides using SVMHC. BMC Bioinformatics, 2002, 3(1): 25

7 Pawlak Z. Rough sets. International Journal of Information and Computer Sciences, 1982, 11: 341～356

8 王国胤. Rough集理论与知识获取. 第1版. 西安交通大学出版社，2001, 17, 118～119

9 王国胤，于洪，杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报，2002, 25(7): 759～766

10 Dan P, Zh Qi-Lun, An Z, H Jing-Song. A novel self-optimizing approach for knowledge acquisition[J]. IEEE Transactions on Systems, Man, and Cybernetics- Part A: Systems and Humans, 2002, 32: 505～514

11 Rammensee H, Bachmann J, Emmerich NP, et al. SYFPEITHI: Database for MHC ligands and peptide motifs[J]. Immunogenetics, 1999, 50: 213～219

12 吴福保，李奇，宋文忠. 基于粗集理论知识表达系统的一种归纳学习方法. 控制理论与决策，1999, 14(3): 206～211

13 Pawlak Z, Slowinski R. Rough set approach to multi-attribute decision analysis. European Journal of Operational Reaserch, 1994, 72: 443～459

(上接第209页)

表3 检测结果比较

算法	离群点数	正确率	误检率	漏检率	运行时间(秒)
SOF	4	100%	0%	0%	4.3
SLOM	3	75%	25%	25%	6.1
SLZ	3	75%	25%	25%	4.1

在SLOF的计算中取$\delta=0.01$，计算结果如表2，如果δ取更小，slof的顺序不变，其值改变，且随着δ变小，彼此的偏差拉大，因此从这种角度看δ越小越好，但也扩大了SLOF的取值范围。

结论 基于SLOF的算法充分考虑了空间数据的特点，根据空间关系确定空间邻居，减少了用户指定参数，用计算邻域距离和空间局部离群系数的方法解决了空间自相关性和异质性问题，并通过对离群系数按降序进行排序，解决了离群点的判断问题。通过与相关算法SLZ和SLOM算法的比较，可以看出SLOF算法在检测精度、计算效率和对用户的依赖性方面均优于其它算法，尤其是在检测的正确性上更为明显。未来通过对SLOF算法的改进，将其应用到基于时间与空间约束的时空离群点的检测中。

参考文献

1 Han Jiawei, Micheline K. Data mining: concepts and techniques [M]. San Francisco: Morgan Kaufmann Publishers, 2001

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

3 Shekhar S, Lu C T, Zhang P. A Unified Approach to Spatial Outliers Detection[J]. GeoInformatica, 2003, 7(2): 139～166

4 Sanjay C, Sun Pei. SLOM: a new measure for local spatial outliers[J]. Knowledge and Information Systems, 2006, 9(4): 412～429

5 Shekhar S, Chawla S. A Tour of Spatial Databases[M]. Upper Saddle River, N. J.: Prentice Hall, 2003

6 Xue Anrong, Ju Shiguang. Algorithm for Spatial Outlier Detection Based on Outlying Degree[C]. In: Proceedings of the WCICA 2006, Dalian: IEEE Press, 12(7): 6005-9