基于熵度量的空间邻域离群点查找

苏锦旗[1],薛惠锋[1],吴慧欣[2]
SU Jin-qi[1],XUE Hui-feng[1],WU Hui-xin[2]

1.西北工业大学 自动化学院,西安 710072
2.华北水利水电学院 信息工程学院,郑州 450011
1.Automation College,Northwestern Polytechnical University,Xi'an 710072,China
2.Dept. of Information Engineering,North China University of Water Conservancy & Electric Power,Zhengzhou 450011,China
E-mail:jinqisu@163.com

SU Jin-qi,XUE Hui-feng,WU Hui-xin.New approach of spatial neighborhood outliers detection based on entropy measurement.Computer Engineering and Applications,2009,45(21):41-43.

Abstract: There are usually two classes of outlier detection algorithms.One is usually applied to statistical data and takes all attributes as multi-dimensional space,while not distinguish between geo-spatial dimensionality and non-spatial dimensionality in detecting process.Meaningless or incorrect outliers can be found if we use these approaches.The other outlier detection algorithms distinguish between geo-spatial dimensionality and non-spatial dimensionality,but they have poor efficiency or can't detect neighborhood outliers.To overcome these shortcomings,new approach of spatial neighborhood outliers detection based on entropy measurement is proposed.In this paper,the spatial attributes are used to determine spatial neighborhood,entropy theory is used to determine the weight of non-spatial attributes,and the non-spatial dimensions are used to compute the spatial neighborhood outlier factor,thus spatial neighborhood outliers can be captured. Theoretical analysis shows that the algorithm is reasonable.The experimental results show that the approach is practical.

Key words: entropy measurement;spatial neighborhood outliers detections;spatial outlier factor;space division

摘 要: 离群点的查找算法主要有两类:第一类是面向统计数据,把各种数据都看成是多维空间,没有区分空间维与非空间维,这类算法可能产生错误的判断或找到的是无意义的离群点;第二类算法面向空间数据,区分空间维与非空间维,但该类算法查找效率太低或不能查找邻域离群点。引入熵权的概念,提出了一种新的基于熵权的空间邻域离群点度量算法。算法面向空间数据,区分空间维与非空间维,利用空间索引划分空间邻域,用非空间属性计算空间偏离因子,由此度量空间邻域的离群点。理论分析表明,该算法是合理的。实验结果表明,算法具有对用户依赖性小、检测精度和计算效率高的优点。

关键词: 熵度量;空间邻域离群点检测;空间邻域偏离因子;空间划分
DOI: 10.3778/j.issn.1002-8331.2009.21.010 **文章编号:** 1002-8331(2009)21-0041-03 **文献标识码:** A **中图分类号:** TP391.9

1 引言

离群点检测在多个领域得到广泛的应用,其任务是从大量复杂的数据中挖掘出存在于小部分异常数据中的新颖的、与常规数据模式显著不同的数据模式[1],研究离群点异常行为有助于发现特别有价值的知识。早期的离群点挖掘算法是针对全部数据集的,挖掘的是全局离群点[2]。在现实世界中对象多为空间对象,应考虑空间对象存在的空间关系,而且在很多场合,用户只关心某些空间邻域的不稳定,这就导致了开展空间邻域离群点的研究。空间领域离群点的检测需要解决邻域的确定和对象与邻域的度量比较两个问题。为此,我们利用空间对象的空间属性及空间关系确定空间邻域,利用空间对象的非空间属性度量对象与其邻域的距离,进而计算每个对象在其邻域的离群度,解决空间邻域离群点的挖掘问题。

2 相关研究

已有的离群点查找方法大多建立在统计学的基础上,大致可以分为以下四类[3]:基于分布的方法,用各种统计模型来测试,将偏离这些模型的对象作为离群点;基于深度的方法,通过计算不同层面的 k-凸面来查找离群点, 凸面的外层认为是离群点;基于距离的离群点,由 Knorr 等提出,若 $P\%$的对象与某对象的距离超过 d,则这个对象为离群点;基于聚类的方法,在查找的过程中,除了聚类剩下的就是离群。

基金项目: 陕西省自然科学基金(the Natural Science Foundation of Shaanxi Province of China under Grant No.2005F45);陕西科技攻关计划(the Key Technologies R&D Program of Shaanxi Province,China under Grant No.2005K04-G13)。

作者简介: 苏锦旗(1979-),男,博士研究生,主要研究方向:人工智能与数据挖掘;薛惠锋(1964-),男,教授,博士生导师,主要研究方向:人工智能、复杂系统建模与仿真;吴慧欣(1978-),男,博士,主要研究方向:计算机图形学、复杂系统建模与仿真。

收稿日期: 2008-04-22 **修回日期:** 2008-06-06

上述方法把多维数据看成同一度量的多维空间,不区分空间维和非空间维,可能导致发现的离群点没有实际意义或产生错误的判断，为此出现专门针对空间离群点的查找方法。Shekhar 等[4]认为空间离群点是空间邻域中非空间属性与其他对象明显不同的空间对象,并提出 SLZ 方法,该方法用空间属性来定义邻域,用非空间属性识别离群点,但它仅适合查找单属性空间离群点。Lu Chang-Tien 等人[5]将此方法进行了拓展(称之为 LCK 方法),可查找多属性的空间离群点。但这些方法仅适合全局离群点。为此,Breunig 等提出了局部离群因子的概念 LOF[6],这是一种基于密度的方法,通过局部的偏离度来判断离群点。LOF 出现后,许多离群度的度量方法相继提出,比较典型的有基于连接的离群系数(COF)[7]、多粒度偏差因子(MDEF)[8]、局部空间离群测度(SLOM)[9]。He 等[10-11]提出了一种基于聚类的局部离群点因子来发现聚类,还提出了分类离群点的概念。基于以上研究,提出了一种基于熵权的空间领域离群点检测方法。

3 非空间属性权重的确定

在计算对象间的 d-维非空间属性距离的时候，不同属性对分析目标的贡献程度不同,这里用熵权作为非空间属性相对重要程度的度量。

定义 1 熵(entropy)是被用来描述数据集合的不确定性(uncertainty),用它来表达正常行为的变动程度。

子空间对象集 $O=\{o_1,o_2,\cdots,o_n\}$中,设有 n 个指标,每个对象有 m-维非空间属性,称之为(m,n)问题。按照定性与定量结合的原则取得对象属性评价矩阵

$$\boldsymbol{R}'=\begin{pmatrix} r'_{11} & \cdots & r'_{1n} \\ \vdots & & \vdots \\ r'_{m1} & \cdots & r'_{mn} \end{pmatrix} \tag{1}$$

对 $\boldsymbol{R}$ 做标准化处理得到:

$$\boldsymbol{R}=(r_{ij})_{m\times n} \tag{2}$$

r_{ij} 称为第 j 个对象在 m-属性上的值,且 $r_{ij}\in[0,1]$,则在 n 个对象 m-维属性中,第 i 维属性的熵定义为:

$$H_i=-k\sum_{j=1}^{n}f_{ij}\ln f_{ij}\quad i=1,2,\cdots,m \tag{3}$$

式中,$f_{ij}=\dfrac{r_{ij}}{\sum_{j=1}^{n}r_{ij}}$,$k=\dfrac{1}{\ln n}$。

为了数据处理方便,可以选择使得 $0\leqslant H_i\leqslant 1$。这样,在$(m,n)$属性度量中,第 i-维属性的熵权 w_i 定义:

$$w_i=\frac{1-H_i}{m-\sum_{i=1}^{m}H_i} \tag{4}$$

式中,$0\leqslant w_i\leqslant 1$,$\sum_{i=1}^{m}w_i=1$,根据熵权确定不同属性对分析目标的贡献程度。

4 空间邻域离群点查找

定义 2 空间关系是指空间对象之间在一定区域内构成的与空间特性有关的联系。

定义 3 给定一个自反的、对称的空间关系 R,空间对象 o_i,o_j 满足空间关系 R,空间对象 o_i 与 o_j 互称为邻居。

定义 4 空间对象 o 的邻域 N_o 是指一个对象集 $O=\{o_1,\cdots,o_m\}$,这里每个 o_i 都是 o 的一个邻居。

定义 5 在对象集 $O=\{o_1,o_2,\cdots,o_n\}$,对象 $o_i,o_j\in O$,o_i 和 o_j 的 M-维非空间属性是 $f(o_i)$和 $f(o_j)$,数据对象 o_i,o_j 之间的 M-维非空间属性加权距离

$$d_A(o_i,o_j,w)=\sqrt{\sum_{i=1}^{M}w_i(f(o_{i_k})-f(o_{j_k}))^2} \tag{5}$$

其中,w_i 为第 i-维属性的熵权,下角 A 代表非空间属性。

定义 6 对象 o 的邻域距离是指对象 o 与其空间邻域中所有对象的加权距离的平均值,即

$$Nd(o,w)=\frac{\sum_{p\in N(o)}d_A(p,o,w)}{|N(o)|} \tag{6}$$

其中,$|N(o)|$为对象 o 空间领域的基。

定义 7(对象领域密度)

$$D(o,w)=\frac{|N(o)|}{\sum_{p\in N(o)}d_A(p,o,w)} \tag{7}$$

定义 8 空间对象 o 所有邻居的邻域密度平均值与 o 的邻域密度的比值得到 o 的空间邻域偏离因子,即

$$K_{SNOF}(o)=\frac{\sum_{p\in N(o)}D(p,w)/D(o,w)}{|N(o)|} \tag{8}$$

$K_{SNOF}(o)$反映对象 o 的非空间属性偏离它周边邻居的程度。

定义 9 给定 n 个对象集 o,希望挖掘 m 个离群点,计算每个对象的 K_{SNOF},K_{SNOF}最大的 m 个对象就是空间离群点。

5 空间偏离度量的修正与构建

空间邻域离群系数反映了空间邻域中对象的非空间属性关系,但这个定义有不足之处。即某些属性值异常的点(如离群点具有极端的属性值),若这些值直接加入计算,将使邻近点的邻域密度不能反映它的真实情况,进而可使它的偏离因子不能反映它的实际偏离程度,例如点 o 的邻域中,由于存在某个极大值(或极小值)会使 $K_{SNOF}(o)$的值变小(或变大),从而不能反映点 o 离群系数的实际意义。为此,设计了一个过滤方法用来过滤这种异常值:在计算邻域空间离群系数的和时,先求已计算的邻域内各个对象与其他对象的属性距离的均值与方差,再用不等式$|(x-\mu)^2/\sigma|<3$ 进行判断，其中满足不等式的被认为是合理的值,被加入属性距离和,而把不符合不等式的值过滤掉,并用平均值代替。

定义 10 修正后对象领域距离

$$\hat{Nd}(o,w)=\frac{\hat{S}_d}{|N(o)|} \tag{9}$$

$\hat{S}_d$ 是修正后的属性距离之和。

定义 11 修正后的空间邻域偏离因子

$$K_{SNOF}(o)=\frac{\sum_{p\in N(o)}\hat{D(p,w)}/D(o,w)}{|N(o)|} \tag{10}$$

这个定义较精确地反映了对象的非空间属性偏离它邻居的程度。

6　基于熵权的空间离群点检测算法

目标:设计一个有效的空间离群点检测算法

条件:获取离群点目标的时间少,准确度高

输入:空间对象集 $O=\{o_1,o_2,\cdots,o_n\}(1\leqslant i\leqslant n)$;

对象 o_i 空间属性集 $s(o_i)$ 和非空间属性集 $f(o_i)$;

空间关系 R;

离群点个数 m;

输出:空间离群点集

算法过程:

(1)根据空间属性和空间关系划分空间邻域;

(2)评价空间对象的非空间属性集,利用熵度量法中式(1)、式(2)、式(3)式(4)求得非空间属性熵权;

(3)根据熵权调整属性指标,利用式(5)计算邻域内对象之间的距离;

(4)剔除邻域中的极值距离,根据式(9)计算修正后对象与邻域平均距离,由此用式(10)计算每个空间对象的空间邻域偏离因子 $\hat{K}_{SNOF}$;

(5)将对象空间偏离因子 $\hat{K}_{SNOF}$ 按降序排序;

(6)输出前 m 个对象,前 m 个对象就是空间离群点;

(7)结合领域知识进行应用分析。

算法的时间复杂性分划分邻域的时间和计算 K_{SNOF} 的时间。由于计算 K_{SNOF} 的 CUP 时间相对是较小的,算法时间复杂性主要决定于邻域计算。计算邻域的时间复杂性与邻域搜索方法有关,这里利用空间索引来划分邻域,其时间复杂性是 $O(knlogn)$。

7　实验

为了验证算法的可行性、准确性和效率,实验采用某市 1:500 地形图,以 280 个小区为研究对象,这些数据包括所有对象的空间和非空间信息,对于每一个小区,与它毗邻的小区组成其邻域,这里选择 5-维非空间属性:Greenrate,Landscape,Facilitation,Security,Perquality,分别对空间属性下的空间离群点算法进行实验分析。

首先通过 10 个评价对象对这 5 个评价标准评判得(5,10)属性矩阵,根据熵度量法,得非空间属性熵权 $P=\{0.18,0.16,0.21,0.19,0.36\}$,利用本文算法(SNOD)挖掘得 5 个空间离群点如表 1 所示。实验结果表明,用这种方法发现的 5 个离群点确实是邻域离群点,并且对小区评价有实际意义。

表 1　SNOD 算法挖掘的 5 个离群点

Num	Outliers	G	L	F	S	P	$K_{SNOF}(o)$	Data
231	YHGY	0.140 2	0.152 8	0.036 44	0.661 6	0.851 1	15.02	Binary
179	HZHA	0.025 5	0.031 2	0.004 9	0.626 3	0.774 1	11.15	Binary
23	HCJY	0.057 6	0.061 1	0.089 5	0.595 8	0.720 8	9.22	Binary
118	QYHY	0.064 1	0.119 4	0.016 9	0.619 5	0.791 5	9.01	Binary
43	JYGJ	0.033 9	0.046 1	0.010 1	0.625 5	0.795 1	8.14	Binary

为了验证算法的性能,利用文献[4]提出的 SLZ 和文献[5]提出的 LCK 查找该空间对象的离群点,查找结果与提出的算法比较如表 2 所示。可以知道,LCK 算法和 SLZ 算法更偏向挖掘全局离群点,而 SNOF 算法更倾向于挖掘邻域离群点。

为了进一步比较算法的效率,对以上实验数据进行人工扩充,随机扩充为 10 000 条数据,20 000 条数据,300 000 条数据,400 000 条数据,50 000 条数据,60 000 条数据,20 个非空间属性做实验。三种方法查找离群点性能比较如图 1 所示。由于三种方法最耗时的是划分空间领域,在耗时上处于同一数量级,但随着数据和维数的增加,SNOD 算法的效率明显好于 SLZ 和 LCK 方法。

表 2　检测结果比较

算法	查找离群点	正确率/(%)
SNOD	YHGY、HZHA、HCJY、QYHY、JYGJ	100
SLZ	YHGY、HCJY、WADS、HZHA、DJGY	60
LCK	YHGY、LMGY、QSC、HCJY、WADS	40

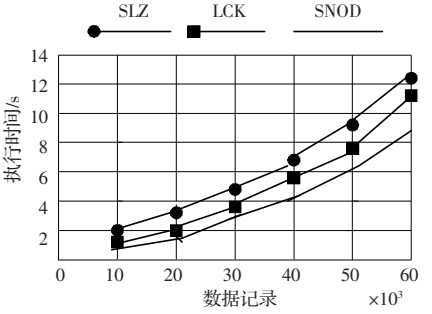

图 1　三种方法查找离群点的效率比较

8　结语

已有的离群点查找方法大多面向统计数据的,造成离群点的遗漏或发现无意义的离群点;新的离群点查找方法将数据属性分为空间属性和非空间属性,利用空间属性建立空间索引,可提高搜索速度。利用非空间属性计算比较离群度,提高算法的可行性。用熵度量法确定空间对象的非空间属性权值,提出了空间偏离因子的度量方法,由此查找空间邻域离群点。实验结果表明,该算法在提高检测精度和降低算法复杂性方面有明显优势。

参考文献:

[1] Han J,Kamber M.Data mining:concepts and techniques[M].San Fransisco,CA,USA:Morgan Kanfmann Publishers,2000:381-389.

[2] HAN Jia-Wei,Micheline K.Data mining:Concepts and techniques[M].2nd ed.San Francisco:Morgan Kaufmann Publishers,2006.

[3] 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.

[4] Shekhar S,Lu Chang-tie,Zhang Pu-sheng.A unified approach to detecting spatial outliers[J].GeoInformatica,2003,7(2):139-166.

[5] Lu Chang-Tien,Chen D-Chang,Kou Yu-Feng.Detecting spatial outliers with multiple attributes[C]//Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence(ICTAI 03),Sacramento,2003:122-128.

[6] Breuning M,Kriegel H P,Ng R T,et al.LOF:Idetifying density-based Local Outliers[C]//Proceedings of ACM SIGMOD Conference,Dallas,Texas,2000:93-104.

[7] Tang J,Chen Z,Fu A,et a1.Enhancing effectiveness of outlier detections for low-density patterns[C]//Proceeding of Advances in Knowledge Discovery and Data Mining 6th PacificAsia Conference,Taipei,China,2002:535-548.

(下转 50 页)