空间孤立点检测*)

文俊浩[1] 吴中福[2] 吴红艳[2]
(重庆大学软件学院 重庆 400030)[1] (重庆大学计算机学院 重庆 400030)[2]

摘 要 空间孤立点是指与邻居具有不连续性的空间点,或者是偏离观测值以至使人们认为是由不同的体系产生的。空间孤立点检测在交通、生态、公共安全、卫生健康、地震、海啸等领域有广泛应用。传统的根据一个非空间属性值进行孤立点判断的方法容易引起孤立点判断失误。作者在针对多个属性进行考虑的基础上,提出以空间维确定邻居关系,非空间维定义距离函数,使用 Mahalanobis 距离检测孤立点,研究一种新的检测空间孤立点的算法,并对时间复杂度进行分析。仿真实验说明算法可以有效地发现大规模空间数据中的孤立点。

关键词 空间孤立点,空间孤立点检测,Mahalanobis 距离,空间数据集

Spatial Outlier Detection Algorithm

WEN Jun-Hao[1] WU Zhong-Fu[2] WU Hong-Yan[2]
(College of Software Engineering, Chongqing University, Chongqing 400030)[1]
(College of Computer Science, Chongqing University, Chongqing 400030)[2]

Abstract Spatial outliers have been informally defined as observations in a data set which appear to be inconsistent with the remainder of that set of spatial data, or which deviate so much from other observations so as to arouse suspicions that they were generated by a different mechanism. Spatial outliers detection is widely used in credit card, fraud detection, public sanitation analysis and tsunami predication and many other fields. Based on multi-attributes, the spatial attributes are used to determine neighborship and non-spatial attributes to define distance function in the paper. A novel algorithm detecting spatial outliers is proposed, which employs Mahalanobis distance to detect outliers. Computation time complexity is analyzed. The simulated experiments demonstrate that our approach can effectively identify local abnormality in large spatial data sets.

Keywords Spatial outliers, Spatial outlier detection, Mahalanobis distance, Spatial data set

1 引言

空间孤立点,即与邻居具有不连续性的空间点,或者是偏离观测值以至使人们认为是由不同的体系产生的。检测空间孤立点在许多地理信息系统和空间数据库中有广泛的应用。这些应用领域包括交通、生态、公共安全、卫生健康、地震、海啸等领域。

空间数据集一般是一些空间参照物体如道路、建筑物城镇等的模型。一般空间物体的属性可以分为两类:空间属性和非空间属性。空间属性包括位置、形状和其他地理上的拓扑关系。非空间属性一般包括长度、高度、建筑时期、名称等。空间物体的邻居关系是基于空间关系(如距离和连接关系)的空间数据集。人们一般比较空间物体时都是比较非空间属性。

空间孤立点是具有与其空间邻居显著不同的非空间属性。空间孤立点可能不止一个属性与空间位置有关。以往的空间孤立点检测是根据某一个非空间属性值进行判断,这样的结果有时候会出现一些判断失误。

作者提出了一种基于空间物体研究多个非空间属性的空间孤立点检测方法,并用仿真实验验证了该方法的可行有效性。

2 相关技术理论

早期的孤立点检测都是基于统计技术的,这些方法可以分为两大类:基于分布的方法和基于深度的方法。基于统计的方法运用一个标准分布去拟合数据集,孤立点是利用概率分布计算得到的[1];基于深度的方法主要依赖于计算几何和k-维凸面的外壳的层次[2]。

随着空间数据挖掘的兴起,Shekhar 等提出了一种用于检测空间孤立点的图元素集[3]。该方法是基于空间点邻居的平均属性值与某一空间点的属性值进行比较,以判断该空间点是否为孤立点的。现已有几种基于统计技术的检测空间孤立点的方法,这些方法可以分为两类:即图论方法和定量测试。图论方法基于可视化的空间数据,这样可以使空间孤立点非常明显[3];定量分析方法从空间数据集中通过计算获得空间孤立点。Scatterplot 和 Moran Scatterplot 是两种典型的代表算法。多维属性检测空间孤立点,传统方法不能使用的一个主要原因是不能使用高维稀疏的空间数据库数据。在高维稀疏空间数据库中甚至每个点都可以看成是孤立点,也可能没有点是孤立点,这是由于所有的数据值都非常相似。实际上,由于欧几里德距离函数采用了所有点的平均值,因此不能反映空间孤立点的特点。降维技术和重新设计距离函数提

*)基金项目:重庆市自然科学基金支持项目(CSTC,2004BB2182)。文俊浩 副教授,博士研究生,主要研究方向:数据挖掘、软件工程。吴中福 教授,博士生导师,主要研究方向:数据挖掘、计算机网络。吴红艳 研究生,主要研究方向:数据挖掘、软件工程。

供了两种新的研究思路。

所有的多属性空间孤立点检测方法是针对非空间属性的。但空间孤立点检测包括两种维:空间维和非空间维。空间维和非空间维应该分开考虑。空间维确定邻居关系,非空间维用来定义距离函数。文中给出的算法是将空间维和非空间维分开单独考虑的。

3 检测算法

下面给出多维属性空间孤立点检测方法以及均值空间孤立点检测算法。

3.1 问题描述

设空间物体 X 由 m 个变量属性值$(y_1,y_2,\cdots,y_m)(m\geqslant 1)$构成,记为 $y=(y_1,y_2,\cdots,y_m)^T$,这里的 T 是转置运算符号。对一个给定的空间点集 $X=(X_1,X_2,\cdots,X_p)(p\geqslant 1)$。属性函数 f 定义为 X 到 R^m 的映射(m 维欧几里得空间),使得每个空间点 X 的函数值 $f(X)$等于属性向量 Y。为方便起见,记为

$$Y_i=f(X_i)=(f_1(X_i),f_2(X_i),\cdots,f_m(X_i))^T$$
$$=(y_{i1},y_{i2},\cdots,y_m)^T,i=1,2,\cdots,p。$$

对给定的整数 k,记 NN_k 为 X_i 的 k 个最近邻居,$(i=1,2,\cdots,p)$。邻居函数 g 定义为 X 到 R^m 的一个映射,使得 $g(X)$的 j 个分量(记为 $g_j(X)$)为所有在 $NN_k(X)$内空间点属性值 y_j 的一个概括统计。

为检测空间孤立点,比较与 X 相邻的 X 中 Y 的所有分量。比较函数和 h 是 f 和 g 的函数,其定义域为 X,值域是 $R^r(r\leqslant m)$。如 $h=f-g$,代表了一种从 X 到 R^m 的映射($r=m$),记为 $h(X_i)$为 h_i。

给定属性函数 f,邻居函数 g 和比较函数 h,如果 h_i 是序列$(h_1,h_2,\cdots,h_n)$的变化显著的点,称 h_i 所对应的点 X_i 为一个空间孤立点。由空间孤立点的定义知道空间孤立点的确定依赖于选取的函数 g 和 h。以下对检测空间孤立点的算法进行规范:

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 $NN_k(X_1),NN_k(X_2),\cdots,NN_k(X_n)$属性函数 f 为:$X\to R^m$;邻居函数 g 为:$X\to R^m$;比较函数 h 为:$X\to R^r$。据此设计一个算法检测空间孤立点。

3.2 空间孤立点检测算法

下面讨论多维属性空间孤立点算法——均值算法。不同的 g 和 h 可能会导致检测出不同的孤立点,选择的标准是检测出绝大部分具有实际意义的孤立点。例如,可以通过导致原因分析检测出的孤立点是否满足要求。

通过计算 $h(X)$与 X 邻居的平均 h 值的 Mahalanobis 距离,考虑属性值的均值和方差以及协方差,找出 f 与 g 之间的不同,即 $h=f-g$,进而可以检测出特殊属性向量。为描述该方法,考虑以下两点:

(1)在特定条件下,可以认为 $h(X)$服从多维正态分布;

(2)如果 $h(X)$的分布函数为 $N_m(\mu,\Sigma)$,即 m 维向量 $h(X)$服从一个均值为 μ,方差-协方差矩阵为 Σ 的多维正态分布,则$(h(X)-\mu)^T\Sigma^{-1}(h(X)-\mu)$服从 χ^2_m 分布,这里 χ^2_m 分布是自由度为 m 的 χ^2 分布。所以 $h(X)$满足$(h(X)-\mu)^T\Sigma^{-1}(h(X)-\mu)>\chi^2_p(\alpha)$的概率为 α,这里的 α 是 χ^2 的分位数。

现假设有 n 个空间物体 $X_1,X_2,\cdots,X_n$ 以及样本 $h(X_1),h(X_2),\cdots,h(X_n)$,计算样本均值 $\mu_s=\frac{1}{n}\sum_{i=1}^{n}h(X_i)$,样本方差-协方差矩阵 $\Sigma_s=\frac{1}{n}\sum_{i=1}^{n}[h(X_i)-\mu_s][h(X_i)-\mu_s]^T$。然后希望 $h(X)$满足$(h(X)-\mu)^T\Sigma^{-1}(h(X)-\mu)>\chi^2_p(\alpha)$的概率为 α。设 $d^2(X)=(h(X)-\mu_s)\Sigma_s^{-1}[h(X)-\mu_s]$,对任意的 X,如果 $d^2(X)$充分大,X 将被认为是空间孤立点。或者说,如果 $d^2(X)>\theta$,则 X 是一个空间孤立点。这里的 θ 是一个依赖于可信度阈值。当然,如果给出的不是空间孤立点检测阈值而是要求检测出的空间孤立点数目 t,这时只需要将 $d^2(X)$按降序排列选取前 t 个即可。

从以上的讨论中可以找出许多探测空间孤立点的算法。选择 g 为属性向量的中心,可得到以下的算法,运用向量中心的原因是样本的中心是一个较好的鲁棒性观测值,称该算法为中心算法。

空间孤立点检测算法(中心算法)算法步骤:

(1)给定空间数据集 $X=(X_1,X_2,\cdots,X_n)$,阈值 θ,属性函数 f 以及最近邻居数 k;

(2)对每个固定 $j(1\leqslant j\leqslant m)$,标准化属性函数 f_j,即 $f_j=\frac{f_j(X_i)-\mu_{f_j}}{\sigma_{f_j}},(i=1,2,\cdots,n)$;

(3)对每个空间点 X_i,计算 k 最近邻居序列 $NN_k(X_i)$;

(4)对每个空间点 X_i,计算邻居函数 g,使得 g_j 等于 $f_j(X)$的中心值,这里 $f_j(X)=\{f_j(X)\mid f_j(X),X\in NN_k(X_i)\}$,比较函数 $h(X_i)=f(X_i)-g(X_i)$;

(5)计算 $d^2(X)=(hX)-\mu_s)\Sigma_s^{-1}[h(X)-\mu_s]$,如果 $d^2(X_i)>\theta$,则 X_i 是空间孤立点。

3.3 计算复杂度分析

中心算法的第 2 步需要标准化分布函数,时间复杂度为 $O(mn)$,第 3 步需要计算每个空间点的 k 个最近邻居(KNN),这里的时间复杂度取决于 KNN 的查询方式,如果采用基于网格的方法且查询网格目录可以装入主内存,此时的时间复杂度为 $O(n)$。如果采用的索引结构(如 R-树),此时的时间复杂度为 $O(\log n)$。在第 4 步中,计算邻居函数 g 和比较函数 h 的时间复杂度为是 $O(mkn)$,在第 5 步中计算 Mahalanobis 距离的时间复杂度是 $O(nm^2)$。总的来说,中心算法的时间复杂度为:当采用网格结构时为 $O(mn)+O(n)+O(mkn)+O(nm^2)$,当采用索引结构时的时间复杂度为 $O(mn)+O(n\log n)+O(mkn)+O(nm^2)$。如果 $n>>k$ 且 $n>>d$,则中心算法的总的时间复杂度采用网格时为 $O(n)$,采用索引时为 $O(n\log n)$。时间复杂度取决于 KNN 查询时采用的策略。

4 仿真实验与分析

数据特征:(1)$n=293$,采用三维空间数据,如图 1 所示。

(2)$m=1$,即每个空间数据点只有一个属性值,便于图形显示,每个数据点的属性值服从均值为 50,标准差为 5 的正态分布,除了 5 个有意设定的孤立点,它们分别是第 50,80,150,210,260 个空间点,这些孤立点处的属性值与其他空间点的分布明显不同。

(下转第 210 页)

参 考 文 献

1 Sutcliffe G, Suttner C. Evaluating general purpose automated theorem proving systems. Artificial Intelligence, 2001, 131(1): 39～54

2 Bray T, Paoli J, Sperberg-McQueen C M. Extensible Markup Language (XML) 1.0, W3C Recommendation, February 1998. http://www.w3.org/ TR/1998/REC-xml-19980210

3 Carlisle D, Ion P, Miner R, et al. Mathematical Markup Language (MathML) Version 2.0, 2001. http://www.w3.org/TR/2001/REC-MathML2-20010221/

4 RuleML, HornML. http://www.dfki.uni-kl.de/ruleml

5 RuleML Initiative. http://www.dfki.de/ruleml

6 王家兵，徐正权，王能超. RLD演绎及子句蕴含与子句包含关系的非等价性. 计算机研究与发展，2002，39(12)：1630～1636

7 Gudgin M, Hadley M, Mendelsohn N, et al. SOAP Version 1.2 Part 1: Messaging Framework, W3C Candidate Recommendation, December 2002. http://www.w3.org/TR/2002/CR-soap12-part1-20021219

8 Gudgin M, Hadley M, Mendelsohn N, et al. SOAP Version 1.2 Part 2: Adjuncts, W3C Candidate Recommendation, Dec. 2002, http://www.w3.org/TR/2002/CR-soap12-part2-20021219

9 Chinnici R, Gudgin M, Moreau J-J, et al. Web Services Description Language (WSDL) Version 1.2, W3C Working Draft, March 2003. http://www.w3.org/TR/2003/WD-wsdl12-20030303

10 Robinson J A. A machine-oriented logic based on the resolution principle. J. ACM, 1965, 12(1): 23～41

11 刘叙华. 基于归结方法的自动推理. 北京：科学出版社，1994

12 Walther C. A Many-Sorted Calculus Based on Resolution and Paramodulation. London: Pitman / Los Alamitos: Morgan Kaufmann, 1987

13 Cohn A G. A more expressive formulation of many sorted logic. J. Automated easoning, 1987, 3: 113～200

14 Beierle C, Hedtstück U, Pletat U, et al. An order-sorted logic for knowledge representation systems. Artificial Intelligence, 1992, 55: 149～191

15 Goguen J A, Meseguer J. Order-sorted algebra I: equational deduction for multiple inheritance, overloading, exception and partial operations. Theoretical Computer Science, 1992, 105: 217～273

16 Weibel T. An order-sorted resolution in theory and practice. Theoretical Computer Science, 1997, 185(2): 393～410

17 Walther C. Many-sorted unification. J. ACM, 1988, 35(1): 1～17

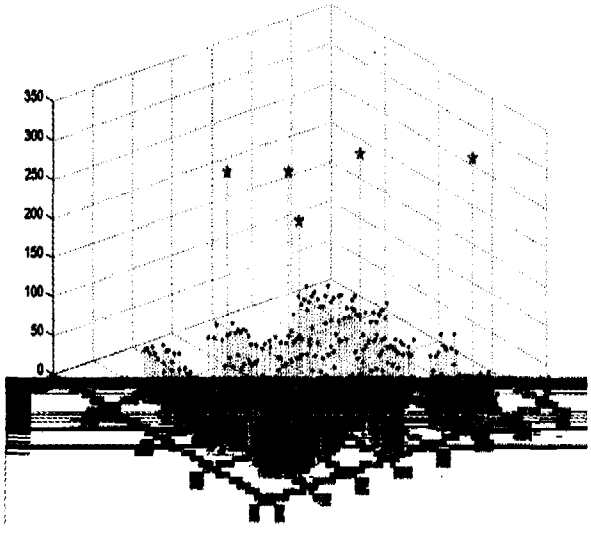
18 Walther C. A mechanical solution of Schubert's steamroller by many-sorted resolution. Artificial Intelligence, 1986, 26(2): 217～224

19 Thompson H S, Beech D, Maloney M, et al. XML Schema Part 1: Structures. W3C Recommendation, May 2001. http://www.w3.org/TR/2001/ REC-xmlschema-1-20010 502/

20 Biron P V, Malhotra A. XML Schema Part 2: Datatypes. W3C Recommendation, May 2001. http://www.w3.org/TR/2001/REC-xmlschema-2-20010502/

21 SOAP Toolkit 3. http://msdn.microsoft.com/xml/, June 2002

(上接第 187 页)



i	50	80	150	210	260
Xi	320	300	280	290	330

图 1 仿真实验的结果

算法在实现时，计算 k 个最近邻居的算法是基于空间数据点的欧氏距离的，其他的均按照上述算法描述实现。运行结果如表 1 所示：将 $d^2(X_i)$ 的值按照降序排列，给出排名前 6 位的情况：

表 1 孤立点检测的运行结果

Xi	X260	X50	X80	X210	X150	X185
$d^2(X_i)$	71.6111	70.1971	51.5573	49.0483	43.1441	7.8590

由上表的结果可知，所检测出的五个孤立点分别为 260，50，80，210，150，与有意设定的五个孤立点相吻合，充分说明了算法的可行有效性。

结束语 文中给出了基于 Mahalanobis 距离的一种空间孤立点检测算法，实验结果表明该方法算法是可行有效的，同时分析了中心算法的时间复杂度。

空间孤立点检测除了上述研究的类型外还有时序数据和时空数据的孤立点检测，这些都涉及到其他的邻居的属性值。后续的工作研究就是进行时空数据的孤立点检测。

参 考 文 献

1 Berchtold S. The pyramid-technique: Towards breaking the curse of dimensionality. In: Proc. ACM SIGMOD Intl. Conf. on Management of Data. ACM Press. 1998. 142～153

2 Luc A. Exploratory Spatial Data Analysis and Geographic Information Systems. In: M. Painho ed. New Tools for Spatial Analysis, 1994. 45～54

3 Shekhar S, Lu C, Zhang P. Detecting Graph-Based Spatial Outlier: Algorithms and Applications (ASummary of Results). In Proc. of the Seventh ACMSIGKDD Int'l Conference on Knowledge Discovery and Data Mining. Aug 2001

4 Shekhar S, Lu C, Zhang P. Detecting Graph-Based Spatial Outlier. Intelligent Data Analysis: An International Journal, 2002. 451～468

5 Panatier Y, Variowin. Software For Spatial Data Analysis in 2D. New York: Springer-Verlag, 1996