

基于空间局部偏离因子的离群点检测算法

张天佑, 王小玲

(中南大学信息科学与工程学院, 长沙 410083)

摘要: 针对空间数据集的特性, 提出一种基于空间局部偏离因子(SLDF)的离群点检测算法。利用 SLDF 度量空间点对象的离群程度, 计算空间数据集中点对象的 SLDF 值并对其进行排序, 将取值较大的前 M 个点对象作为空间离群点。实验结果表明, 该算法能较好地检测空间局部离群点, 其有效性与准确性均优于 SLZ 算法, 适用于高维大数据集的空间离群点检测。

关键词: 属性权重向量; 空间离群点; 空间对象距离; 空间局部偏离因子

Outlier Detection Algorithm Based on Space Local Deviation Factor

ZHANG Tian-you, WANG Xiao-ling

(College of Information Science and Engineering, Central South University, Changsha 410083, China)

【Abstract】 According to the characteristics of spatial data sets, this paper proposes an outlier detection algorithm based on the Space Local Deviation Factor(SLDF). The algorithm uses SLDF to measure the deviate degree of space points object. It calculates all the points' SLDF, sorts by their values, and uses the top M as the space outlier. Experimental result shows that the algorithm can well detect space outlier and be more applicable to the high dimensional and large data sets, its validity and accuracy of the algorithm are superior to that of SLZ algorithm.

【Key words】 attribute weighted vector; space outlier; space object distance; Space Local Deviation Factor(SLDF)

DOI: 10.3969/j.issn.1000-3428.2011.14.096

1 概述

离群点检测是数据挖掘中的重要研究领域之一, 常用于异常检测、商业欺诈行为监测、网络入侵检查等方面。离群点检测在国外得到十分广泛的研究与应用, 并引起越来越多学者与专家的重视。文献[1]给出离群点的定义: 离群点的表现与其他点不同, 它可能是由另外一种完全不同的机制产生的。目前, 离群点的挖掘算法主要有以下 4 类: 基于分布的, 基于距离^[2]的, 基于密度^[3]的和基于深度的算法, 每种算法都给出了对离群点的定义。早期离群点检测算法一方面是针对全部数据集的, 检查出的离群点是全局离群点; 另一方面由于对各项参数的设置比较敏感, 经常会出现漏检与误检的情况, 因此检测的有效性与精度都不高。

本文主要是对空间离群点挖掘算法进行研究, 对空间离群点的度量方式进行进一步改进。针对空间数据集的特性, 本文充分考虑空间点对象的属性权重(属性权重可以事先由专家给定, 也可以通过其他方式计算得到, 关于如何得到属性权重值并不是本文研究的重点), 以点对象非空间属性带权重距离作为空间点对象距离, 给出一种新的空间离群点度量标准: 空间局部偏离因子(Space Local Deviation Factor, SLDF), 从而更好地解决空间离群点的挖掘问题。

2 基于空间局部偏离因子的离群点检测算法

2.1 相关概念与定义

定义 1(属性权重向量) 数据集 $X = \{X_1, X_2, \dots, X_n\}$, 其中, $X_i = (x_{i1}, x_{i2}, \dots, x_{id})$ 为第 i 个数据对象; 属性集 $A = \{A_1, A_2, \dots, A_d\}$; 权重向量 $w = (w_1, w_2, \dots, w_d)$; A_i 对应的权值为 w_i , $\sum_{i=1}^d w_i = 1$ 。

定义 2(空间邻居) 对象 O 的空间邻居^[4]是指对象 O 在指定条件 C 下, 存在空间邻接关系 sc 的对象。

定义 3(空间 k 距离邻域) 空间对象 O 的空间 k 距离邻域是指与对象 O 的带权重距离小于 k 的所有空间邻居集合, 即 $\forall o \in O, N(k, o) = \{dist(p, o, w) < k\}$, 其中, k 表示给定对象 o 的距离。

定义 4(空间对象距离) 设 $X_i, X_j \in X$, w_k 是第 k 维的权值, $0 \leq w_k \leq 1$, 对象 X_i 和 X_j 之间的距离定义为:

$$dist(X_i, X_j, w) = \sqrt{\sum_{k=1}^d w_k [f(x_{ik}) - f(x_{jk})]^2}$$

其中, $\sum_{k=1}^d w_k = 1$ 。需要指出的是, 这里对象距离并非对象间的空间距离(空间距离表示方式有欧式距离、曼哈顿距离等多种表示方法), 而是对象间基于 d 维非空间属性的加权距离。

定义 5(对象 p 的空间局部偏离率) 设对象 p 的空间 k 距离邻域为 $N_{k-distance}(p)$, 简记为 $N_k(p)$, $|N_{k-distance}(p)|$ 表示邻域的基, 即邻域内所有邻居的个数; 对象 p 到其邻域内所有邻居的平均值记为 μ , 即: $\mu = \frac{\sum_{o \in N_k(p)} dist(p, o, w)}{|N_k(p)|}$ 。

定义对象 p 的空间局部偏离率为:

$$S_{SLDR_k}(p) = \frac{\sum_{o \in N_k(p)} [dist(p, o, w) - \mu]^2}{|N_k(p)|}$$

其中, 分子是空间对象 p 与其邻域内邻居距离的偏差; 分母是对象 p 的空间 k 距离邻域邻居的总数。

对象 p 的空间局部偏离率反映了 p 的空间 k 距离邻域内

基金项目: 国家自然科学基金资助项目(60773013)

作者简介: 张天佑(1983—), 男, 硕士研究生, 主研方向: 数据挖掘; 王小玲, 教授

收稿日期: 2011-02-16 **E-mail:** 285264628@qq.com

数据集对对象 p 的影响。如果 $S_{SLDR_k}(p)$ 的值很小, 说明对象 p 的周围的对象分布比较均匀, 则对象 p 成为离群点的可能性较小。如果 $S_{SLDR_k}(p)$ 的值很大, 说明对象 p 的周围的对象分布比较稀疏, 则对象 p 成为离群点的概率较大。

定义 6(对象 p 的空间局部偏离影响率) 给定对象 p 的空间 k 距离邻域为 $N_{k-distance}(p)$ 、对象 p 的空间局部偏离率 $S_{SLDR_k}(p)$, 定义对象 p 的空间局部偏离影响率为:

$$S_{SLDIR_k}(p) = \frac{\sum_{o \in N_k(p)} S_{SLDR_k}(o)}{|N_k(p)|}$$

其中, 分子是对象 p 的空间 k 距离邻域邻居集中对象的 SLDR 之和。对象 p 的空间局部偏离影响率反映了 p 的空间 k 距离邻域内数据集对对象 p 的偏离影响程度

定义 7(对象 p 的空间局部偏离因子) 给定对象 p 的空间局部偏离率 $S_{SLDR_k}(p)$ 、 p 的空间局部偏离影响率 $S_{SLDIR_k}(p)$, 定义 p 的空间局部偏离因子为:

$$S_{SLDF_k}(p) = \frac{S_{SLDR_k}(p)}{S_{SLDIR_k}(p)}$$

对象 p 的空间局部偏离因子反映了 p 的空间 k 距离邻域内邻居对象的分布情况。如果 SLDF 值比较大, 说明对象 p 的周围空间是一个稀疏的区域, 该对象很可能成为一个离群点; 如果 SLDF 值比较小, 说明该对象周围空间是一个稠密区域, 它不可能成为一个离群点。

定义 8(空间离群点) 给定数据集 $X = \{X_1, X_2, L, X_n\}$, 其中, $X_i = (x_{i1}, x_{i2}, L, x_{id})$ 表示第 i 个数据对象, d 表示维数。若挖掘 m 个离群点, 分别计算每个对象的 SLDF 值, 并按其大小排序, 则 SLDF 值最大的前 m 个对象即为空间离群点^[5]。

2.2 本文算法实现步骤与描述

本文算法具体步骤如下:

输入 数据集 $X = \{X_1, X_2, L, X_n\}$, 记 $S(X_i)$ 、 $F(X_i)$ 分别表示其空间属性与非空间属性。 D 维非空间属性 $F(X_i)$ 表示为 $(f(x_{i1}), f(x_{i2}), L, f(x_{id}))$; 属性集 $A = \{A_1, A_2, L, A_d\}$; 权向量 $\omega = (\omega_1, \omega_2, L, \omega_d)$; A_i 对应的权值为 ω_i , $\sum_{i=1}^d \omega_i = 1$, 正整数 k 为空间距离邻域参数值, m 为预期挖掘空间离群点个数。

输出 空间离群点集

(1) 归一化每个对象的非空间属性值;

For each object x_i in DataSet (X) Do //对数据集中每个对象 //进行归一化处理

For each dimension j in AttributeSet (A) Do

Calculate $f(x_{ij})$ //其中, $f(x_{ij}) = (f(x_{ij}) - \min(f(x_{ij}))) / (\max(f(x_{ij})) - \min(f(x_{ij})))$

(2) 根据非空间属性值与权向量, 计算数据集中任意 2 个对象之间的空间对象距离;

For each object x_i in DataSet (X) Do

For each object x_j in DataSet (X) Do

For each w_k in w Do

Calculate $\text{dist}(x_i, x_j, w)$ by definition 4 //其中, $\text{dist}(x_i, x_j, w)$ 为 //对象 X_i 与 X_j 的空间对象距离(即非空间属性带权距离)

(3) 根据给定参数 k , 从数据集中随机选取一个对象作为起始点, 依次计算每个对象的空间邻域;

While($\text{dist}(x_i, x_j, w) \leq k$)

Add x_j to $N(k, x_i)$ //其中, $N(k, x_i)$ 为对象 X_i 空间 k 距离邻居 //集合

(4) 依次计算数据集中每个给定对象的空间局部偏离率;

For each object x_i in DataSet (X) Do

Calculate each object x_i SLDR value by definition 5

(5) 通过剪枝过程(剪掉空间局部偏离率为零的对象), 得到离群点的候选集;

If(SLDR!=0)

Add object x_i to CandidateOutlierSet

Else Delete object x_i in DataSet(X)

(6) 依次计算候选集中每个给定对象的空间局部偏离影响率;

For each object x_i in CandidateOutlierSet Do

Calculate each object x_i SLDIR value by definition 6

(7) 依次计算候选集中每个给定对象的空间局部偏离因子;

For each object x_i in CandidateOutlierSet Do

Calculate each object p SLDF value by definition 7

(8) 按 SLDF 值大小对其进行降序排列;

Sort all object order by SLDF value

(9) 根据给定参数 m , 取前 m 个对象作为结果进行输出。

Output the top_m_object which have higher SLDF value;

2.3 算法时间复杂度分析

本文算法执行所需时间由以下 6 个步骤构成:

步骤 1 归一化空间点集的非空间属性值, 时间复杂度为 $O(d \times n)$ 。

步骤 2 求数据集任意 2 点间的空间对象距离, 其时间复杂度为 $O(d \times n \times n)$ 。

步骤 3 计算空间点对象 k 距离邻域, 时间复杂度为 $O(n)$ 。

步骤 4 根据候选离群点集计算点对象的 SLDF 值, 时间复杂度为 $O(k \times r)$ 。

步骤 5 排序, 时间复杂度为 $O(m \times r)$ 。

其中, m 为前 m 个离群点的个数; k 为输入参数值; r 为候选离群点集规模; d 为属性集维数; n 为数据集对象的个数(即数据集的基(|X|))。

3 实验结果与分析

实验计算机硬件运行环境: CPU 为 Intel P4 1.52 GHz, 内存为 256 MB, 外存 40 GB; 软件运行环境: 操作系统为 Windows XP, 算法在 Matlab7.1 下实现。

3.1 数据描述

为验证本文算法的有效性与准确性, 本实验所用仿真数据集为二维空间数据集。为了便于图形显示与实验结果观察, 模拟数据集利用程序在长为 10 cm、宽为 8 cm 的二维空间中产生 5 000 个数据点(数据记录号从 1~5 000)及 100 个随机分布于整个区域的离群点(数据记录号从 1~100)。本实验空间邻居数 k 设置为 60, 挖掘的离群点数 m 设置为 40。

3.2 实验分析

本实验主要利用本文算法与 SLZ 算法^[6]进行比较。本文算法检测出的离群点示意图如图 1 所示, SLZ 算法检测出的离群点示意图如图 2 所示。实验结果表明, SLZ 算法能识别的离群点, 本文算法基本都能识别。从图 1 可以看出, 本文算法能够识别出的离群点更多, 原因在于 SLZ 算法是基于全局离群点的识别, 而本文算法能更有效地发现空间局部离群点。以点对象空间局部偏离因子进行离群程度的度量很好地体现了局部性, 从而具有更高的检测精度。



图 1 SLZ 算法检查出的离群点结果

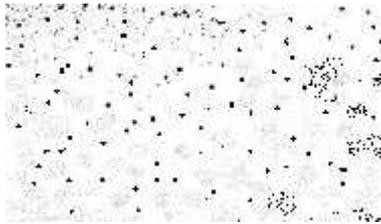


图2 SLDF算法检查出的离群点结果比较

本文算法与SLZ算法检测的前6个离群点对比情况见表1。为进一步测试本文算法的有效性,在上述实验样本数据集的基础上,将数据集依次扩充到6 000个、8 000个、10 000个记录,离群点个数(100个)与其他参数保持不变。分别用2种算法对其进行检测,其检测所需时间与检测出离群点情况如表2所示。

表1 2种算法检测出的前6个离群点对比情况

顺序	本文算法		SLZ算法	
	记录号	SLDF值	记录号	SLZ值
1	45	9.893	67	32.042
2	78	6.792	56	29.359
3	67	5.017	93	16.783
4	32	3.169	69	14.594
5	14	1.082	38	11.643
6	93	0.534	16	8.772

表2 2种算法的检测性能比较

算法	数据集大小/个	执行时间/s	漏检数	误检数
本文算法	6 000	23	12	15
	8 000	38	9	12
	10 000	49	4	7
SLZ算法	6 000	28	21	23
	8 000	55	14	17
	10 000	63	10	11

(上接第281页)

可以看到,对于固定模型参数的方法,随着时间的增加,模型精度逐渐恶化。而采用动态更新模型系数的方法,模型的精度逐渐提高,最终稳定在-42 dB左右。这说明功放的特性在较长的时间尺度上缓慢的波动和变化,而要获得好的DPD性能,需要能够动态实时地提取模型参数,以适应功放的缓变特性。

与动态模型相比较,传统的LSM算法计算所得的模型NMSE为-43.1 dB,较动态模型好大约1 dB。前文中的各个模型的功率谱密度图也给出类似的结论,这是由于传统LSM算法通常选用的数据点要比动态算法的数据窗口更长,因此更不容易收到各种噪声和测量误差的影响,但这是以牺牲计算时间和效率为前提所获得的。实际上,可以看到这2种模型的差别很小,在绝大多数应用场景中完全可以忽略,而采用动态模型大幅度提高系统的效率和实时性。

6 结束语

本文介绍的动态模型能够更准确地描述宽带功率放大器的特性,并且能够快速收敛到较高的模型精度,满足宽带

实验结果表明,在相同数据集下,本文算法检测效率与准确性更高,并且更适用于大数据集的空间离群点检测。

4 结束语

本文提出一种基于空间局部偏离因子的离群点检测算法。针对空间数据集的特殊性,该算法不仅考虑了空间点对象的属性权重,还提出一种新的空间离群点度量方式,用数据点的空间局部偏离因子作为度量空间点对象的离群程度。在二维仿真数据集上进行实验,并与文献[6]的SLZ算法进行比较。理论分析与实验结果表明,本文算法能够更好地检测出空间局部离群点,其检测的有效性与准确性均优于本文算法,并且更适用于高维大数据集上的空间离群点检测。今后工作是将本文算法运用于真实数据集的空间离群点检测。

参考文献

- [1] Hawkins D. Identification of Outliers[M]. London, UK: Chapman and Hall, 1980.
- [2] Zoubi M B A, Obeid N. A Fast Distance Algorithm to Detect Outliers[J]. Journal of Computer Science, 2007, 3(12): 944-947.
- [3] Breunig M M, Kriegel H P, Ng R T, et al. LOF: Identifying Density-based Local Outliers[C]//Proceedings of International Conference on Management of Data. [S. l.]: ACM Press, 2000.
- [4] 王妍, 潘瑜春, 阎波杰. 基于Voronoi和空间自相关的离群点检测[J]. 计算机工程, 2010, 36(1): 33-35.
- [5] 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.
- [6] Shekhar S, Lu Changtie, Zhang Pusheng. A Unified Approach to Detecting Spatial Outliers[J]. GeoInformatica, 2003, 7(2): 139-166.

编辑 陆燕菲

DPD研究的要求,可以用于实时性要求较高的DPD应用场景,同时也为进一步展开宽带DPD技术的研究提供参考。

参考文献

- [1] Pedro J C, Maas S. A Comparative Overview of Microwave and Wireless Power Amplifier Behavioral Modeling Approaches[J]. IEEE Trans. on Microwave Technology, 2005, 53(3): 1150-1163.
- [2] Wood J. Fundamentals of Nonlinear Behavioral Modeling for RF and Microwave Design[M]. Norwood, UK: [s. n.], 2005.
- [3] Pedro J C. Pruning the Volterra Series for Behavioral Modeling of Power Amplifiers Using Physical Knowledge[J]. IEEE Trans. on Microwave Technology, 2007, 55(5): 813-821.
- [4] Nelles O. Nonlinear System Identification[M]. Berlin, Germany: Springer-Verlag, 2001.
- [5] 曹新容, 黄联芬, 赵毅峰. 一种最小二乘/奇异值分解算法[J]. 计算机工程, 2009, 35(16): 278-280.

编辑 陈文