第33卷 第23期 计 算 机 工 程 2007年12月
Vol.33 No.23 Computer Engineering December 2007
·博士论文· 文章编号：1000—3428(2007)23—0035—02 文献标识码：A 中图分类号：TP311

基于 Voronoi 图的异常检测算法

曲吉林[1]，寇纪淞[2]，李敏强[2]，安世虎[1]
(1. 山东财政学院计算机与信息工程学院，济南 250014；2. 天津大学系统工程研究所，天津 300072)

摘 要：异常检测是数据挖掘的一个重要组成部分，其中基于密度的方法LOF是目前常用的主要方法。然而LOF方法进行检测时需要设定参数*k*和*MinPts*，检测结果对参数非常敏感，容易造成检测错误。该文提出了一种基于Voronoi图的异常检测算法VOD，采用Voronoi图来确定对象间的邻近关系，解决了基于密度方法存在的问题，算法的时间复杂性从$O(N^2)$降低到$O(N\log N)$。

关键词：数据挖掘；异常检测；基于密度；Voronoi图

Outlier Detection Algorithm Based on Voronoi Diagram

QU Ji-lin[1], KOU Ji-song[2], LI Min-qiang[2], AN Shi-hu[1]
(1. School of Computer and Information Engineering, Shandong University of Finance, Jinan 250014;
2. Institute of System Engineering, Tianjin University, Tianjin 300072)

【Abstract】 Outlier detection is an integral part of data mining, and the density-based method LOF is the current state of the art in outlier detection. However, LOF is very sensitive to its parameter *k* and *MinPts*, which may result in wrong estimation. This paper proposes a new outlier detection algorithm based on Voronoi diagram called VOD. VOD measures the outlier factor automatically by Voronoi neighborhoods without parameter, which provides highly-accurate outlier detection and reduces the time complexity from $O(N^2)$ to $O(N\log N)$.

【Key words】 data mining; outlier detection; density-based; Voronoi diagram

1 概述

异常检测(outlier detection)是数据挖掘的一个重要组成部分。近年来，国内外学者提出了一系列异常检测方法，包括基于统计的方法、基于距离的方法、基于密度的方法、基于聚类的方法和基于偏差的方法等，其中基于密度的方法[1]由于检测性能和效率比较高，成为当前异常检测的主要方法。

基于密度的异常检测方法的基本思想是：根据对象 p 在其 k-邻域内的局部密度，与 k-邻域内的其它点相比较，计算其局部异常因子(local outlier factor, LOF)，来描述一个对象的异常程度。这种方法克服了基于距离的方法中不同密度子集混合造成的检测错误，但其仍然存在以下问题：

(1)算法的时间复杂性比较高，达到$O(dN^2)$。其中，d为数据的维数；N为数据集中对象的个数。

(2)计算对象p的局部密度需要确定参数k和$MinPts$，不仅需要用户具有相关领域的先验知识，而且检测结果对k和$MinPts$的选择非常敏感，容易造成检测结果错误。

Jin等分析了参数k对异常检测结果的影响[2]。在图 1 所示的点集中，聚集c_1的密度高于c_2的密度，p点的异常因子显然低于点q，但LOF方法计算出p点的异常因子反而比q点的异常因子高，这显然是错误的。

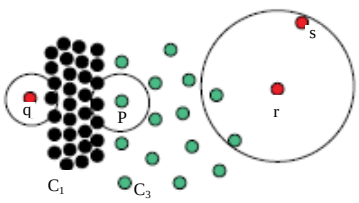

图 1 参数k对异常检测结果的影响

针对基于密度方法存在的问题，近年来国内外学者进行了深入的研究，提出了许多改进方法[2~4]。这些方法在一定程度上改进了LOF的性能，但只是对LOF的局部优化。正如文献[3]所说，目前仍然没有一种简单有效的方法解决基于密度方法存在的问题。Pei等[4]指出，基于密度和距离的异常检测问题的瓶颈是邻近点的搜索。

本文采用Voronoi图来描述点的邻近关系，定义了一种新的V邻域异常因子(Voronoi neighbor outlier factor, VNOF)，提出了一种基于Voronoi图的异常检测算法(Voronoi outlier dection, VOD)，不需要预先设置k和$MinPts$等参数，解决了基于密度方法存在的缺陷。一方面避免了参数选择带来的各种问题，另一方面将算法的时间复杂性从$O(N^2)$降低到$O(N\log N)$。

2 VOD 算法的基本原理

2.1 Voronoi 图的相关知识

Voronoi 图如图 2 所示。

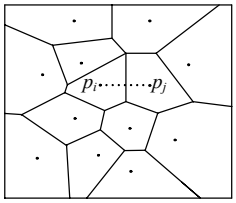

图 2 Voronoi 图

作者简介：曲吉林(1963－)，男，教授、博士，主研方向：数据挖掘；寇纪淞、李敏强，教授、博士生导师；安世虎，教授、博士

收稿日期：2007-08-01 **E-mail：**qujl@sdfi.edu.cn

设p_i和p_j是平面上的两个点，线段p_ip_j的垂直平分线将平面分为两部分，用$H(p_i, p_j)$表示包含p_i的半平面，$H(p_j, p_i)$表示包含p_j的半平面。显然，$H(p_i, p_j)$内的点比平面上其它点更接近p_i，即$H(p_i, p_j)$由到p_i的更近的点构成，记作$V(p_i)=H(p_i, p_j)$；同理，$V(p_j)=H(p_j, p_i)$表示平面上其它点更接近p_j的点。

定义 1 给定N个点的集合

$$S=\{p_1,p_2,\cdots,p_N\},V(p_i)=\bigcap_{i\neq j}H(p_i,p_j)$$

则$V(p_i)$表示比其他点更接近p_i的N-1 个半平面的交，它是一个不多于N-1 条边的凸多边形区域，$V(p_i)$称为关联于p_i的Voronoi多边形；点集S所有点的Voronoi多边形将平面划分为N个区域，称为点集S的Voronoi图，记作Vor(S)。Voronoi图的点称为Voronoi顶点，线段称为Voronoi边。

根据Voronoi多边形的定义，给定$p_i \in S$，$V(p_i)$包含且只包含S中的一个点p_i。因此，对于$V(p_i)$多边形区域内的任意一点，其到p_i的距离比到S中其它点的距离都小。

Voronoi 图在理论和实际中具有广泛的应用，这里只讨论与本文有关的性质。

定理 1 在点集S中，p_i的每一个邻近点确定Voronoi多边形$V(p_i)$的一条边[5]。

上述定理说明如何根据p_i的邻近点确定$V(p_i)$的边。类似地，通过$V(p_i)$的边可以找出p_i的所有邻近点，记作$V_N(p_i)$。

定理 2 在点集S中，p_i的Voronoi多边形$V(p_i)$的每一条边确定p_i的一个邻近点，并且可以确定p_i的所有邻近点[5]。

定理 3 N个点的Voronoi图至多有 2N-5 个顶点和 3N-6 条边[5]。

定理 4 对N个点的点集S，在$O(N\log N)$时间内能够构造出S的Voronoi图，这在时间上是最优的[5]。

构造点集S的Voronoi图的主要方法有分治法和平面扫描法等。

定理 5 利用定点集S的Voronoi图，能够在$O(\log N)$时间内找出某点的所有邻近点，这在时间上是最优的[5]。

给定点集S，构造点集S的Voronoi图。根据定理 2，与p_i关联的Voronoi多边形$V(p_i)$的每一条边就确定p_i的一个邻近点，如图 2 所示。

2.2 VOD 算法的原理

根据上述讨论，对点集S中的一点p_i，通过p_i的Voronoi多边形$V(p_i)$来确定其邻近点，计算p_i的局部密度和异常因子，这样不仅更加准确、合理，同时也避免了参数设置引起的检测结果错误。

定义 2 对点集S的任意一点p，由$V(p)$边确定的p的邻近点称为p的V-邻近点，p所有V-邻近点的集合记作$V_N(p)$。

定义 3 点p所有V-邻近点到p的平均距离的倒数，称为p的V-近邻分布密度，记作$V_d(p)$。即

$$V_d(p)=1\Big/\left(\sum_{o\in V_N(p)}d(p,o)\Big/|V_N(p)|\right) \tag{1}$$

其中，$|V_N(p)|$为p所有V-邻近点的个数。$V_d(p)$反映了点p周围点的分布密度。对于异常点，与其V-邻近点的平均距离比较大，其近邻分布密度相应地比较小。

定义4 点p的V-邻域异常因子VNOF定义为

$$\mathrm{VNOF}(p)=\frac{1}{|V_N(p)|}\sum_{o\in V_N(p)}\frac{V_d(o)}{V_d(p)} \tag{2}$$

从上述定义可以看出，利用点p近邻点的分布密度与该点分布密度的比值来确定异常点。对于具有不同分布密度的数据集，异常点的邻近点的分布密度与其自身的分布密度的比值应比其他非异常点要大。

3 VOD 算法的描述与分析

算法 基于Voronoi图的异常检测算法VOD。

输入 点集S，异常点数λ。

输出 点集S中各点的V-邻域异常因子和异常点。

(1)构造点集S的Voronoi图Vor(S)。

(2)对S的每个点p_i，计算其V-近邻分布密度$V_d(p_i)$。

(3) 对点集S的每个点p_i，计算其V-邻域异常因子VNOF(p_i)。

(4)根据VNOF(p_i)值从大到小排序。

(5)输出各点的V-邻域异常因子，以及异常因子最大的前λ个点。

算法中，λ为预计的异常点数，由用户根据相关领域知识确定。实际应用中，也可以将λ设为异常因子的阈值，这样第(5) 步只要输出VNOF(p_i)>λ的点即可，也可以直接根据VNOF(p_i)值的大小由用户确定异常点。设点集S中包含N个点，对于算法的时间复杂性分析如下：

算法第(1)步构造点集S的Voronoi图，根据定理 4，时间为$O(N\log N)$，这是算法的预处理时间。

第(2)步计算每个点的V-近邻分布密度，关键是找出每个点的V-邻近点。对于点p_i，根据定理 2，其Voronoi多边形$V(p_i)$的一条边确定p_i的一个V-邻近点，p_i的V-邻近点的个数等于多边形$V(p_i)$的边数。Voronoi图的每条边由相邻的 2 个点所共有，计算所有点的V-邻近点的次数等于Voronoi图边数的 2 倍。根据定理 3，N个点的Voronoi图至多 3N-6 条边，因此，找出每个点的V-邻近点的时间不超过 2(3N-6)，即算法第(2)步的时间为$O(N)$；同理，算法第(3)步的时间也是$O(N)$。

算法第(4)步根据VNOF(p_i)值从大到小排序，时间为$O(N\log N)$。

算法第(5)步输出各点的V-邻域异常因子，以及异常因子最大的前λ个点，时间为$O(N)$。

根据上述分析，整个算法的时间复杂性为$O(N\log N)$。

4 实验

为了测试 VOD 异常检测方法的性能，分别用图 1 中的点集和实际数据，对 VOD 和 LOF 方法进行了实验，LOF 方法中取 k=3。实验环境为 P4 630 3.0GHz，Windows XP 操作系统，采用 MATLAB 和 C 语言编程。

实验 1 采用图 1 中的数据。LOF 方法的检测结果为：LOF(p)= 1.2245, LOF(q)= 0.9778。

本文提出的 VOD 方法的检测结果为：VNOF (p)= 1.1407, VNOF (q)= 1.1614。

基于密度的 LOF 方法中，p 点的异常因子高于 q 点的异常因子，检测结果显然是错误的，而 VOD 方法给出了正确的结果。

实验 2 采用 1980 年 1 月 1 日至 1992 年 10 月 8 日 IBM 公司股票每日收盘价时间序列，共 3 333 个数据。数据来自 http://www-personal.buseco.monash.edu.au/~hyndman/TSDL/korsan/dailyibm.dat。

对股票收盘价时间序列分段线性化，得到收盘价连续上升或下降的持续时间和斜率，用点(持续时间，斜率)表示，

（下转第 39 页）