一种基于 Voronoi 图的高效异常检测方法

曲吉林
QU Ji-lin

山东财政学院 计算机与信息工程学院,济南 250014
Department of Computer and Information Engineering,Shandong University of Finance,Ji'nan 250014,China
E-mail:qujl@sdfi.edu.cn

QU Ji-lin.Efficient outlier detection method based on Voronoi diagram.Computer Engineering and Applications,2008,44(3):178-179.

Abstract: Outlier detection is an integral part of data mining.In this paper,we propose a new outlier detection method based on Voronoi diagram.The new method 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;Voronoi diagram

摘 要:提出了一种新的基于 Voronoi 图的异常检测方法。采用 Voronoi 图来确定对象间的邻近关系,定义了一种新的异常因子,算法的时间复杂性为 $O(n\log n)$。实验结果表明,同现有的算法相比具有较高的检测效率和准确性。
关键词:数据挖掘;异常检测;Voronoi 图
文章编号:1002-8331(2008)03-0178-02 **文献标识码**:A **中图分类号**:TP391

1 引言

异常检测(Outlier Detection)是数据挖掘的一个重要组成部分,在入侵检测、金融、电信、气象和医疗等领域都具有广泛的应用。

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。Hautamäki 等在文献[1]中提出了一种基于 K-邻近图的方法,用 K-邻近点的平均距离来衡量各点的异常程度,具有较高的检测准确度,但算法的时间复杂性为 $O(n^2)$,检测效率不高,同时参数 k 的选择需要用户具有相关领域的先验知识。

本文采用 Voronoi 图来描述点的邻近关系,提出了一种新的基于 Voronoi 图的异常检测方法,不需要预先设置 k 等参数,同时将算法的时间复杂性从 $O(n^2)$降低到 $O(n\log n)$。

2 基本原理

2.1 Voronoi 图的相关知识

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,如图 1 所示。

定义 1 Voronoi 图。给定 n 个点的集合 $S=\{p_1,p_2,\cdots,p_n\}$,令 $V(p_i)=\bigcap_{i\neq j}H(p_i,p_j)$,则 $V(p_i)$表示比其它点更接近 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。

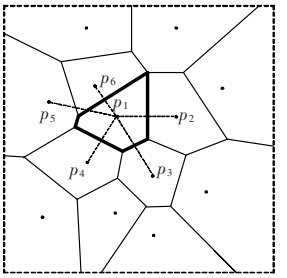


图 1 Voronoi 图

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。

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[2]。

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 pi 的所有邻近点。如图 1,由 $V(p_1)$确定的 p_1 的邻近点为 p_2、p_3、p_4、p_5 和 p_6。

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[2]。

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

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[2]。

基金项目:山东省科技攻关计划(the Key Technologies R&D Program of Shandong Province,China under Grant No.2007GG3WZ10010);山东财政学院博士科研启动基金资助(No.06BSJJ09)。
作者简介:曲吉林(1963-),男,博士,教授,主要研究领域为数据挖掘、计算几何等。

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[2]。

给定点集 S,首先构造点集 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,这样就可以找出 p_i 的所有邻近点。

2.2 V-异常因子

从 Voronoi 图的定义和性质可以看出,对于点集 S 中不同的点,其邻近点的数量不一定相同,这是由点集的分布决定的。选择一个固定的整数 k,通过 p_i 的 k 个邻近点来确定 p_i 的异常程度,是不合理的。

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 的到其各邻近点的平均距离,用平均距离的倒数来衡量 p_i 的异常程度。

定义 2 V-邻近点。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 $V(p)$。

定义 3 V-异常因子。点 p 所有 V-邻近点到 p 的平均距离的倒数,称为 p 点的 V-异常因子,记作 $V_d(p)$。即:

$$V_d(p)=\frac{1}{\left(\sum_{o\in V(p)}\frac{d(p,o)}{|V(p)|}\right)}$$

其中$|V(p)|$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。$V_d(p)$越大,表明 p 点周围点集的分布越稀疏,其异常因子也就越大。

3 算法描述与分析

算法 基于 Voronoi 图的异常检测方法

输入:点集 S,异常点数 λ

输出:点集 S 中各点的 V-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-异常因子 $V_d(p_i)$。

(3)根据 $V_d(p_i)$从大到小排序。

(4)输出各点的 V-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。实际应用中,也可以将 λ 设为异常因子的阈值,这样第 4 步只要输出 $V_d(p_i)>\lambda$ 的点即可。

设点集 S 中包含 n 个点,对于算法的时间复杂性分析如下:

算法步骤(1)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。

步骤(2)计算各点的 V-异常因子 $V_d(p_i)$,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,因此 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 $O(n)$。

算法步骤 3 根据各点的 V-异常因子 $V_d(p_i)$ 从大到小排序,时间为 $O(n\log n)$。

算法步骤 4 输出各点的 V-异常因子,以及异常因子最大的前 λ 个点,时间为 $O(n)$。

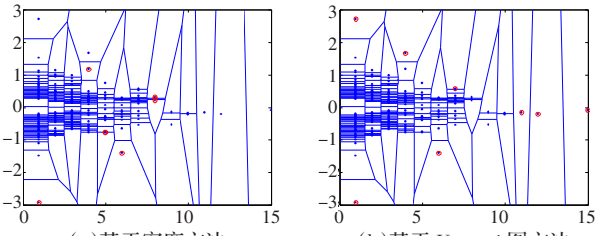
根据上述分析,整个算法的时间复杂性为 $O(n\log n)$。

4 实验结果

为了验证算法的准确性,采用实际数据,利用 MATLAB 和 C 语言编程,对本文提出了方法和目前最常用的基于密度的方法[3]进行了对比实验。

实验数据采用迪士尼公司 1996 年 3 月 29 日~1999 年 3 月 29 日股票的每日收盘价,共 756 个数据[4]。

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,分别利用三种方法分析股票的异常波动。基于密度的方法中取参数 k=3,设定前 8 个异常因子最大的点为异常点,用圆圈标出,实验结果如图 2 所示。



(a)基于密度方法检测到的异常点　(b)基于 Voronoi 图方法检测到的异常点

图 2 实验结果

从图 2 中可以看出,基于密度方法检测出的 4 个异常点在密度较高的区域中,而提出的基于 Voronoi 图方法给出了正确的结果。

5 结束语

本文利用 Voronoi 图的基本原理,定义了一种新的 V-邻域异常因子,提出了一种新的基于 Voronoi 图的异常检测方法,同现有的方法相比具有较高的准确度和效率。

对于高维空间中的异常检测,受"维灾"问题的影响,随着数据维数的升高,算法的效率急剧下降,目前主要采用近似算法。文献[5]提出了一种高维空间中近似 Voronoi 图的构建方法,同其它高维空间中的异常检测相比,算法同样具有较高的检测效率。(收稿日期:2007 年 8 月)

参考文献:

[1] Hautamäki V,Kärkkäinen I,Fränti P.Outlier detection using k-nearest neighbour graph[C]//Proceedings of the 17th International Conference on Pattern Recognition,Cambridge,UK,2004,3:430-433.

[2] Preparata F P,Shamos M I.Computational geometry-An introduction[M].New York:Springer-Verlag,1985.

[3] Breunig M M,Kriegel H P,Ng R,et al.LOF:Identifying Density-based Local Outliers [C]//Proceedings of the ACM SIGMOD International Conference on Management of Data,Dallas,Texas,USA,2000,93-104.

[4] MathWorks Inc.Financial time series toolbox[EB/OL].[2007-05].http://www.mathworks.com.

[5] Har P S.A replacement for voronoi diagrams of near linear size[C]//Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science,Las Vegas,Nevada,USA,2001:94-103.