计算机研究与发展 ISSN 1000-1239/CN 11-1777/TP
Journal of Computer Research and Development 47(12): 2110－2116, 2010

一种基于密度的局部离群点检测算法 DLOF

胡彩平 秦小麟
(南京航空航天大学信息科学与技术学院 南京 210016)
(hucaiping@nuaa.edu.cn)

A Density-Based Local Outlier Detecting Algorithm

Hu Caiping and Qin Xiaolin
(*College of Information Science and Technology, Nanjing University of Aeronautics & Astronautics, Nanjing* 210016)

Abstract With rapid growth of data, data mining becomes more and more important. Detecting outlier is one of the very important data mining techniques, which is to find exceptional objects that deviate from the most rest of the data set. There are two kinds of outliers: global outliers and local outliers. In many scenarios, the detection of local outliers is more valuable than that of global outliers. The LOF algorithm is a very distinguished local outlier detecting algorithm, which assigns each object an outlier-degree value. However, when the outlier-degree value is calculated, the algorithm should equally consider all attributes. In fact, different attributes have different effects. The attributes with more large effects are known as outlier attributes. In this paper, a density-based local outlier detecting algorithm (DLOF) is proposed, which educes outlier attributes of each data object by information entropy. The weighted distance is introduced to calculate the distance of two data object, which those outlier attributes are assigned with bigger weight. So the algorithm improves outlier detection accuracy. In addition, when the local outlier factors are calculated, we present our two improvements of the algorithm and their time complexity analysis. Theoretical analysis and experimental results show that DLOF is efficient and effective.

Key words local outlier; density; local outlier factor; information entropy; outlier attribute

摘 要 离群点可分为全局离群点和局部离群点.在很多情况下,局部离群点的挖掘比全局离群点的挖掘更有意义.提出了一种基于密度的局部离群点检测算法 DLOF.该方法通过引入信息熵用于确定各对象的离群属性,在计算各对象之间的距离时采用加权距离,并给离群属性较大的权重,从而提高离群点检测的准确度.另外,该算法在计算离群因子时,采用了两步优化技术,并对采用这两步优化技术后算法的时间复杂度进行了详细分析.理论分析和实验结果表明了该方法是有效可行的.

关键词 局部离群点;密度;局部离群因子;信息熵;离群属性

中图法分类号 TP301.6;TP18

0 引 言

离群点检测是数据挖掘领域一个重要的研究方向,用于发现数据集中的噪声数据.离群点就是在数据集中与其他数据点表现不一致的对象或者大大地偏离其他数据点以至于怀疑它是由不同的机制生成的对象.离群点检测可以应用到许多领域,如网络入

收稿日期:2009-07-20;修回日期:2010-07-08
基金项目:国家"八六三"高技术研究发展计划基金项目(NO2007AA01Z404);国家自然科学基金项目(60673127);南京航空航天大学科研启动基金项目(S0848-042);南京航空航天大学基本科研业务费专项科研基金项目(NS2010094)

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

侵检测、电信和信用卡欺骗、气象预报、客户分类等. 现有的发现离群点的方法大多建立在统计学的基础上, 可分为以下几类:

1) 基于分布的[1]. 基于分布的方法在统计领域较为常见, 人们用各种统计模型来测试, 把偏离这些模型的对象当作离群点. 但是, 大多数分布模型只能直接应用于单变量的特征空间, 难以应用于多维空间. 而且, 这种模型要求预先知道数据的分布, 但这种知识往往难以获得, 这就要进行耗时的测试来决定.

2) 基于深度的[2]. 基于深度的方法是一种基于计算几何的方法, 这种方法计算不同层面的 k-凸面来查找离群点, 凸面的外层被认为是离群点. 这个算法对二维和三维空间上的数据比较有效, 但对四维及四维以上的数据, 处理效率比较低.

3) 基于距离的[3-6]. Knorr 与 Ng[3] 提出了一个基于距离的离群点概念. 在数据库中 p%的对象与某对象的距离超过 d, 这个对象就是离群点. 文献[4] 对基于距离的孤立点的概念进行了扩展, 根据 k-最近邻距离对离群点进行排序, 并给出了一种有效的方法, 计算排在最前面的 n 个离群点. 文献[5] 针对不确定数据, 提出基于距离的不确定数据离群点检测的高效过滤方法, 文献[6] 提出了基于 CD-Tree 的离群点检测算法.

4) 基于密度的[7-12]. Breunig 等人[7] 提出了局部离群因子的概念 LOF, 这是一种基于密度的方法. 通过数据空间的所有维度来计算对象的距离, 进而计算对象的可达密度, 最后通过局部的离群度来判断离群点. 该算法不是去界定哪些数据对象是离群点, 哪些数据对象不是离群点, 而是转向量化地描述数据对象的离群程度, 通过赋予每一个数据对象一个表征它离群程度的量化指标来进行离群点的搜索. 自从 LOF 算法出现后, 出现了许多离群度的度量方法, 比较典型的有基于局部信息熵的加权子空间离群点检测算法[8]、基于连接的离群系数[9]、多粒度偏差因子[10] 和局部空间离群测度[11-12] 等方法.

借鉴 LOF 算法的特点, 本文提出了一种基于密度的局部离群点检测算法 DLOF. 该方法通过引入信息熵用于确定各对象的离群属性, 在计算各对象之间的距离时采用加权距离, 并给离群属性较大的权重, 从而提高离群点检测的精度. 另外, 该算法在计算离群因子时, 采用了两步优化技术, 提高了算法的执行效率.

1 LOF 算法的基本概念

LOF 算法的核心思想, 就是通过赋予每一个数据对象一个表征该数据对象偏离程度的因子, 而不是明确地来界定哪些数据对象是离群点而哪些数据对象不是离群点, 而这个表征一个数据对象偏离程度的数值, 实质上则反映了该数据对象是否分布在数据对象较为集中的局部区域之中. 下面介绍本文中需要用到 LOF 算法的一些基本概念[7].

定义 **1**. 对象 p 的第 k 距离(k-distance of an object p).

对于一个正整数 k, 数据对象 p 的第 k 距离记作 $k\text{-}distance(p)$. 在数据集 D 中, 存在 1 个数据对象 o, 该数据对象与数据对象 p 之间的距离记作 $d(p, o)$.

满足以下条件, 则取 $k\text{-}distance(p)$ 等于 $d(p, o)$:

1) 至少存在 k 个数据对象 $o' \in D \setminus \{p\}$ 满足 $d(p, o') \leqslant d(p, o)$;

2) 至多存在 $k-1$ 个数据对象 $o' \in D \setminus \{p\}$ 满足 $d(p, o') < d(p, o)$.

由于每一个数据对象所处的局部空间区域的数据对象分布情况都有所差异, 所以不能够以一个全局化的统一标准来考察每一个数据对象. 该定义通过考察每一个数据对象与被考察数据对象之间的距离并找出其中数值上为第 k 大的那个距离, 来确定一个针对该数据对象的个性化的局部空间区域的范围, 对于数据对象密度较大的区域, 该数值一般情况下较小; 对于数据对象密度较小的区域, 该数值一般情况下则较大.

定义 **2**. ε 邻域(ε-neighborhood).

设 p 为数据集 D 中的一个数据对象, 则数据对象 p 的 ε 邻域记作:

$$N_\varepsilon(p) = \{x \in D \mid d(p, x) \leqslant \varepsilon\}.$$

定义 **3**. 数据对象 p 的第 k 距离邻域(k-distance neighborhood of an object p).

已知数据对象 p 的第 k 距离, 那么, 数据对象 p 的第 k 距离邻域则是所有到 p 的距离小于等于 p 的第 k 距离的数据对象的集合, 记作:

$$N_{k\text{-}dis(p)}(p) = \{q \in D \setminus \{p\} \mid d(p, q) \leqslant k\text{-}dis(p)\}.$$

在本文中, 将 $k\text{-}distance(p)$ 简写为 $k\text{-}dis(p)$, $N_{k\text{-}dis(p)}(p)$ 简写为 $N_k(p)$.

该定义是以被考察对象为圆心, 该数据对象的

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第 k 距离为半径的区域内的所有数据对象的集合(不包括该对象本身). 由于数据集中可能同时存在多个第 k 距离的数据对象, 因此该集合至少包含了 k 个数据对象. 一般情况下, 每一个数据对象的该集合所包含的数据对象个数都不会比 k 大太多, 而只是每一个数据对象的该集合所涵盖的区域存在的差异性较大. 可以想象: 偏离程度较高的数据对象, 该集合所涵盖的区域则较大; 而偏离程度较低的数据对象, 该集合所涵盖的区域则较小. 对于处于同一个聚类群组中的数据对象来说, 它们涵盖的区域面积则大致相当.

定义 **4**. 数据对象 p 相对于数据对象 o 的可达距离(reachability distance of object p w. r. t object o).

设 k 为一自然数, 则数据对象 p 相对于数据对象 o 的可达距离记作:

$$reach\text{-}dis_k(p,o)=\max\{k\text{-}dis(o),d(p,o)\}.$$

定义 **5**. 局部可达密度(local reachability density of an object p).

设 $MinPts$ 为一正整数, 则局部可达密度记作:

$$lrd_{MinPts}(p)=1\Bigg/\left(\frac{\sum_{o\in N_{MinPts(p)}} reach\text{-}dis_{MinPts}(p,o)}{|N_{MinPts}(p)|}\right).$$

在该定义中, 首先计算了所考察数据对象 p 的第 k 距离邻域内的所有数据对象到数据对象 p 的可达距离之和, 从这里我们可以看出, 如果数据对象 p 是一个偏离程度非常高的数据对象, 则它的第 k 邻域所涵盖的范围非常之广, 因为在它周围的空间区域中的邻近数据对象非常少, 而对于 p 的第 k 邻域中的大部分数据对象而言, 因为 p 的偏离程度很高, 因此数据对象 p 到它们的 $reach\text{-}dis_k(p,o)$ 取 $d(p,o)$ 的可能性则较大. 实质上, 已经可以看出, 若某一个数据对象 p 的偏离程度很大, 则对在该数据对象的第 k 距离邻域内的数据对象 o 而言, 数据对象 p 不在 o 的第 k 邻域内的概率就很大, 反之则较小; 若某一个数据对象 p 处于某一个聚类群组之中, 则对在该数据对象的第 k 邻域内的对象 o 而言, p 在 o 的第 k 邻域内的概率就很大, 因此 $reach\text{-}dis_k(p,o)$ 取第 k 距离的概率则较大, 这样导致在某一个群组之内的所有数据对象的 $lrd_{MinPts}(p)$ 数值非常接近, 而偏离程度较大的数据对象的该数值则较小并且与群组中的数据对象之间形成较大的差异性, 通过这样的方式, $lrd_{MinPts}(p)$ 表征了数据对象所处局部空间区域的密度情况.

定义 **6**. 局部离群因子(local outlier factor of an object p).

$$LOF_{MinPts}(p)=\frac{\sum_{o\in N_{MinPts}(p)}\frac{lrd_{MinPts}(o)}{lrd_{MinPts}(p)}}{|N_{MinPts}(p)|},$$

分析该式可知, $\frac{\sum_{o\in N_{MinPts}(p)} lrd_{MinPts}(o)}{|N_{MinPts}(p)|}$ 是计算出数据对象 p 的第 k 距离邻域内的所有数据对象的局部区域密度的平均值, 该数值反映了在数据对象 p 的第 k 距离范围之内的空间点的平均分布密度, 若对象 p 是偏离程度较大的数据对象, 则它的第 k 距离邻域内的数据对象则大多数是距离 p 较远并且处于某一个群组中的数据对象, 那么这些数据对象的 lrd 数值则较大, 该式的计算结果也较大, 而 p 本身由于是偏离程度较大的数据对象, 其 lrd 数值则较小, 最终导致 LOF 数值较大. 反之, 若数据对象 p 处于某一个群组之中, 则其第 k 距离邻域内的数据对象与其属于同一个群组的可能性较大, 所得出的平均分布密度与该数据对象的 lrd 数值差异不会很大, 其 LOF 数值接近于 1, 这已经成功剥离了不同群组之间的密度差异所带来的影响.

2 基于密度的局部离群点检测 DLOF 算法

2.1 DLOF 算法的相关概念

为了提高离群点的检测的质量, 在计算局部离群因子时, 对数据集中两数据对象距离的计算采用加权距离. 本文中采用信息熵来确定加权距离的权重.

熵是信息论中用来描述信息和随机变量不确定性的重要工具, 设 X 为随机变量, 其取值集合为 $S(X)$, $P(x)$ 表示 X 可能取值的概率, 则 X 的熵定义为

$$E(X)=-\sum_{x\in S(X)}P(x)\,\mathrm{lb}\,P(x).$$

变量的不确定性越大, 熵也就越大, 把它搞清楚所需要的信息量也就越大; 熵值越小, 不确定性越小.

假设 d 维数据集 D 的属性集为 $A=\{A_1,\cdots,A_d\}$, D 中数据点 p 在属性 A_i 上的值记为 $f_{A_i}(p)$. 以下是 DLOF 算法的相关定义.

定义 **7**. 离群属性.

对于 $p\in D$, $A_i\in A$, 若满足以下关系:

$$E(A_i)\geqslant\frac{\sum_i E(A_i)}{d},$$

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

则称属性 A_i 为离群属性.

由定义 7 可知, 若属性 A_i 的信息熵不小于其余属性信息熵的均值, 则说明属性 A_i 表现出更大的不确定性. 从信息熵角度考虑, 离群点的存在使得数据集的不确定性增强, 而离群点所表现出的不确定性正是通过其在某些属性上的取值分布造成的. 因此, 把满足该条件的属性称为离群属性.

定义 **8**. 加权距离.

设 $p, q \in D$, 其中 $f_{A_i}(p)$和 $f_{A_i}(q)$是第 i ($i=1, 2, \cdots, d$)维属性的值, w_k是第 k 维的权值, 则数据对象 p 和 q 之间的加权距离为

$$d(p, q, w) = \sqrt{\sum_{i=1}^{d} w_k \ (f_{A_i}(p) - f_{A_i}(q))^2},$$

其中, $w_i = \begin{cases} \lambda, & \lambda > 1 \text{ 如果 } A_i \text{ 是离群属性,} \\ 1, & \text{如果 } A_i \text{ 不是离群属性.} \end{cases}$

2.2 DLOF 算法描述

输入: d 维数据集 D, k, λ, 离群因子阈值 ξ;

输出: 数据集 D 的离群点集合.

算法过程:

1) 计算各个属性的信息熵, 如果某个属性的信息熵大于所有属性信息熵的平均值, 则该属性为离群属性;

2) 计算各对象之间的加权距离, 进一步计算出各对象局部可达密度;

3) 运用定义 6 中公式计算各对象的局部离群因子 LOF;

4) 如果某个对象的离群因子 LOF 大于离群因子阈值 ξ, 则该对象为离群点.

2.3 DLOF 算法的复杂度分析

查找离群点, 就是要利用定义 6 中公式计算每个数据对象的离群因子. 设空间点的总数为 N, 定义 6 中公式可以改写为

$$LOF_{MinPts}(p) = \frac{\sum_{o \in N_{MinPts}(p)} lrd_{MinPts}(o)}{|N_{MinPts}(p)| \times lrd_{MinPts}(p)}.$$

由此可以看出算法 DLOF 的复杂度分为 3 个部分之和. 第 1 部分为 lrd 的复杂度, 即计算对象 p 的局部可达密度; 第 2 部分为计算 p 的第 $MinPts$ 距离邻域的复杂度, 第 3 部分即遍历 p 的第 $MinPts$ 距离邻域的复杂度(基本算术运算和集合大小的计算均为常数时间).

$$O(LOF_{MinPts}) = O(lrd_{MinPts}) + O(N_{MinPts}) + O(|N_{MinPts}| \times O(lrd_{MinPts})).$$

在计算对象 p 的 LOF 数值时, 需要首先计算出 p 的若干个性化的参数, 包括 $N_{MinPts}(P)$, $MinPts\text{-}dis(p)$, $lrd_{MinPts}(p)$. 而在 $lrd_{MinPts}(p)$的计算过程中, 需要计算 $reach\text{-}dis_k(p, o)$, 由定义可知:

$$reach\text{-}dis_k(p, o) = \max\{k\text{-}dis(o), d(p, o)\}.$$

那么, 若在对 $N_{MinPts}(P)$中的某一个对象 o 进行其 LOF 数值的计算时, 若对象 p 也在 o 的 $N_{MinPts}(O)$ 集合之中, 则会重复计算 $MinPts\text{-}dis(p)$ 以及 $d(p, o)$, 由此可知, 若存在对象 p 和 q, 满足 $p \in N_{MinPts}(q)$ 以及 $q \in N_{MinPts}(p)$, 则这两个对象的 $reach\text{-}dis$ 以及 $d(p, q)$均被重复计算了一次, 而对于一个密度较大的群组中的点, 对于每一个点 o 的集合 $N_{MinPts}(o)$中的点集合 N_{MinPts}包含点 o 的可能性非常大, 若点 o 的集合 $N_{MinPts}(o)$之中存在 m 个满足这样条件的点, 则 $MinPts\text{-}dis(o)$和两点之间的距离会被重复计算 m 次.

同样的问题出现在第 k 距离邻域的计算过程之中, 在 LOF 的计算过程之中, 需要计算 $N_{MinPts}(p)$, 而对于 $N_{MinPts}(p)$中的每一个对象 o, 在它的 lrd 计算过程中需要计算 $N_{MinPts}(o)$, 因此可以考虑通过适当增大空间开销来获取时间效率的提高.

第 1 步优化: 在算法执行之前, 先计算所有对象的 $k\text{-}dis(p)$值并保存起来以作备用. 在算法的执行过程之中凡是需要用到 $k\text{-}dis(p)$的地方皆直接使用预先计算好的资源. 然后, 计算所有对象 p 的集合 $N_{MinPts}(p)$并保存起来以备用, 这一步的优化可降低算法时间复杂度的阶.

在该优化中, 算法需要增加规模为 N 的空间开销来存储 $k\text{-}dis(p)$的数值, 而对于空间的第 k 距离邻域, 算法则需要增加规模为 N^2 级的空间开销, 以一定的空间来换取时间的方式有效地降低了算法执行时间. 实验表明, 该改进将至少 10 倍以上地提高算法执行的时间效率, 效率的提升随着计算规模的扩大而增大.

第 2 步优化: 在第 1 步优化的基础之上, 在算法执行之前, 先计算所有对象两两之间的距离并保存起来以作备用. 在算法的执行过程之中凡是需要用到两点之间距离的地方皆直接使用预先计算好的资源. 这一步的优化可以降低算法时间复杂度的隐含常数.

第 2 步的优化是在第 1 步优化的基础之上将两两空间点之间的距离预先计算并存储, 该步计算会进行 $\frac{1}{2}n(n-1)$次循环, 因为两点之间的距离计算在被认为在常数时间内完成, 因此增加的时间开销

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

为 $O(N^2)$, 这种改进不会增加算法的时间复杂度的阶. 实验表明, 该步改进可以将算法的执行时间在第 1 步的基础之上再加快 1 倍左右. 而付出的代价就是需要增加至少 $\frac{1}{2}n(n-1)$, 即 N^2 规模的空间开销.

3 实验结果及分析

本节通过实验对 DLOF 算法的检测精确度和效率进行分析. 测试的硬件环境是: CPU2.5 GHz 的, 主存为 512MB. 软件环境是: 操作系统为 Microsoft Windows XP Professional, 该实验程序使用 C ♯编写, 采用的开发环境为 Microsoft Visual Studio 2005. 测试所用的数据为模拟数据集 Test Dataset 和文献[8]中所使用的测试数据集 KDD CUP1999, 数据集 KDD CUP1999 是网络入侵检测数据集, 该数据集中的数据对象分为五大类, 包括正常的连接、各种入侵和攻击等. 选择了其中的 20 个属性, 对非数值属性进行数值化处理, 对连续属性进行离散化处理.

1) DLOF 算法的优化测试

通过不同规模的数据和参数值对 DLOF 算法和它的两个改进方法进行测试, 以比较其执行时间效率. 表 1 是在参数 k 为 1 情况下的测试结果, 该情况下大部分数据对象的邻域集合的元素个数为 1, 为算法的最好情况. 由表 1 可见, 虽然 3 种方法的时间复杂度下界均相同, 但是改进 2 的实际运行时间基本上是改进 1 的一半, 而改进 1 和改进 2 都比原始方法有了实质性的提高. 所以, 在实际的应用之中, 可以综合考虑硬件条件和时间要求以及所面临的问题的规模, 选择比较折中的计算方法, 以满足应用的需求. 另外, 我们选取数据对象为 900 个的情况下, 测试参数 k 对算法的影响(如图 1 所示). 从图 1 中可以看出随着 k 的增加, 算法的运行时间都有所增加, 但改进方法增加较慢.

Table 1 Runtime Comparison of Different Datasets

表 1 几种数据规模下的运行时间 ms

The Number of Objects	k	DLOF	The First Improvement	The Second Improvement
100	1	62	31	15
200	1	234	62	46
300	1	531	125	78
400	1	937	234	109
500	1	1453	359	171
600	1	2187	515	296
700	1	3046	718	406
800	1	4046	937	484
900	1	5140	1218	656

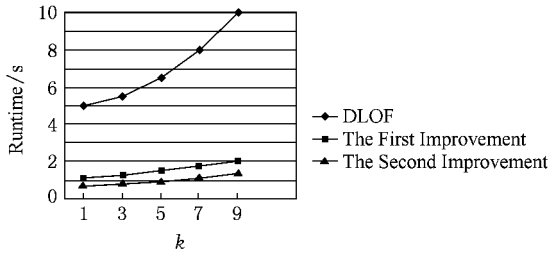

Fig. 1 Runtime comparison of DLOF and two improvements.

图 1 3 种方法的运行时间比较

2) 算法的精确度和执行效率

为了测试算法的精确度和执行效率, 对算法 DLOF, SPOD 和 LOF 的性能进行对比, 采用 KDD CUP1999 数据集和模拟数据集 Test Dataset 作为测试集, 为了进行实验, 对 KDD CUP1999 数据适当的修改, 使得攻击(即离群点)占数据集的 5%. 评价一个离群点检测方法的好坏, 可以通过在给定的数据集上来运行该方法, 并且计算在由该方法所找出的离群点中, 真正的离群点所占据的比例. 比例越高, 则表明该方法的性能越好. 精确度采用以下量度对算法进行评价:

$$\text{精确度}=\frac{\text{正确找到的离群点数}}{\text{离群点总数}}.$$

实验结果如图 2 和图 3 所示, 从图可以看出, DLOF 算法的时间效率与 LOF 算法相近, 由于 DLOF 算法需要计算信息熵来确定离群属性, 所以 LOF 算法的执行速度比 DLOF 算法略快. 但如果采用 DLOF 算法的两个改进方法将比 LOF 算法好很多. 另外, DLOF 算法和 LOF 算法的时间效率比 SPOD 算法都高, 因为 SPOD 算法要计算局部信息熵, 需要花费较多时间. SPOD 算法在精度上比 DLOF 算法和 LOF 算法都要好, 但 DLOF 算法和 SPOD 算法之间相差不大.

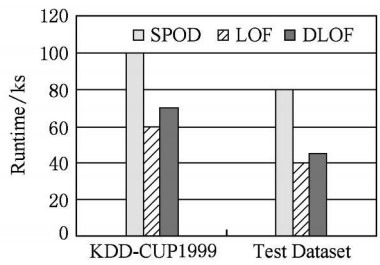

Fig. 2 Runtime comparison of three methods.

图 2 3 种算法执行效率对比

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

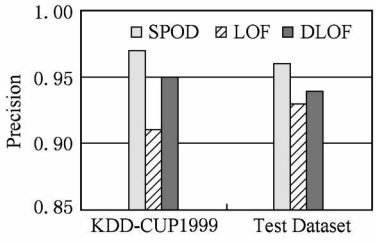

Fig. 3 Precisions of three methods.
图 3 3种算法精确度对比

4 结束语

离群点检测是数据挖掘中一个非常重要的研究领域,具有广泛的应用领域.本文提出了一种基于密度的局部离群点检测算法 DLOF.该方法通过引入信息熵用于确定各对象的离群属性,在计算各对象之间的距离时采用加权距离,并给离群属性较大的权重,从而提高离群点检测的精度.另外,该算法在计算离群因子时,采用了两步优化技术,提高了算法的执行效率.未来我们将考虑进行不确定数据离群点检测和复杂数据对象(包括空间数据、时空数据以及多媒体数据等)的离群点检测.

参 考 文 献

[1] Barnett V, Lewis T. Outliers in Statistical Data[M]. New York: John Wiley and Sons, 1994

[2] Johnson T, Kwok I, Ng R T. Fast computation of 2-dimensional depth contours[C] //Proc of the 4th Int Conf on Knowledge Discovery and Data Mining (KDD'98). New York: ACM, 1998: 224-228

[3] Knorr E M, Ng R T. Algorithms for mining distance-based outliers in large datasets[C] //Proc of the 24th Int Conf on Very Large Data Bases. New York: ACM, 1998: 392-403

[4] Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for mining outliers from large data sets[C] //Proc of the 2000 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2000: 93-104

[5] Yu Hao, Wang Bin, Xiao Gang, et al. Distance-based outlier detection on uncertain data [J]. Journal of Computer Research Development, 2010, 47(3): 474-484 (in Chinese)
(于浩, 王斌, 肖刚, 等. 基于距离的不确定离群点检测[J]. 计算机研究与发展, 2010, 47(3): 474-484)

[6] Sun Huanliang, Bao Yubin, Yu Ge, et al. An algorithm based on partition for outlier detection [J]. Journal of Software, 2006, 17(5): 1009-1016 (in Chinese)
(孙焕良, 鲍玉斌, 于戈, 等. 一种基于划分的孤立点检测算法[J]. 软件学报, 2006, 17(5): 1009-1016)

[7] Breunig M M, Kriegel H P, Ng R T, et al. LOF: Identifying density-based local outliers [C] //Proc of ACM SIGMOD Conf. New York: ACM, 2000: 427-438

[8] Ni Weiwei, Chen Geng, Lu Jieping, et al. Local entropy based weighted subspace outlier mining algorithms [J]. Journal of Computer Research Development, 2008, 45(7): 1189-1194 (in Chinese)
(倪巍伟, 陈耿, 陆介平, 等. 基于局部信息熵的加权子空间离群点检测算法[J]. 计算机研究与发展, 2008, 45(7): 1189-1194)

[9] Tang J, Chen Z, Fu A, et al. Enhancing effectiveness of outlier detections for low-density patterns [C] //Proc of Advances in Knowledge Discovery and Data Mining 6th Pacific Asia Conf. Berlin: Springer, 2002: 535-548

[10] Papadimitirou S, Kitagawa H, Gibbons P B, et al. LOCI: Fast outlier detection using the local correlation integral [C] //Proc of the 19th Int Conf on Data Engineering. Los Alamitos: IEEE Computer Society, 2003: 315-326

[11] Sanjay C, Pei Sun. SLOM: A new measure for local spatial outliers [J]. Knowledge and Information Systems, 2006, 9(4): 412-429

[12] Xue Anrong, Ju Shiguang, He Weihua, et al. Study on algorithms for local outlier detection[J]. Chinese Journal of Computers, 2007, 30(8): 1454-1463 (in Chinese)
(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)

Hu Caiping, born in 1977. PhD and lecturer. His main research interests include spatial data mining and spatial database.
胡彩平, 1977 年生, 博士, 讲师, 主要研究方向为空间数据挖掘和空间数据库等.

Qin Xiaolin, born in 1953. Professor and PhD supervisor. His main research interests include spatial database, spatial data mining and information security.
秦小麟, 1953 年生, 教授, 博士生导师, 主要研究方向为空间数据库、空间数据挖掘和信息安全等.

Research Background

With the rapid growth of data, data mining becomes more and more important. Detecting outlier is one of the very

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

important data mining techniques. There are two kinds of outliers: global outliers and local outliers. In many scenarios, the detection of local outliers is more valuable than that of global outliers. The LOF algorithm is a very distinguished local outlier detecting method, which assigns each object an outlier-degree value. In this paper, a density-based local outlier detecting algorithm (DLOF) is proposed, which educes outlier attributes of each data object by information entropy. The weighted distance is introduced to calculate the distance of two data objects, and those outlier attributes are assigned with bigger weight. So the algorithm improves outlier detection accuracy. In addition, when the local outlier factors are calculated, we present our two improvements of the algorithm and their time complexity analysis. This research is supported by the National High-Tech Research and Development Plan of China (NO2007AA01Z404), the National Natural Science Foundation of China (60673127), NUAA Research Funding(S0848-042, NS2010094).

《计算机应用》征订启事

《计算机应用》于 1981 年创刊,中国计算机学会会刊,由中国科学院成都计算机应用研究所和四川计算机学会主办,科学出版社出版.

《计算机应用》系中文核心期刊、中国科技核心期刊.被《中国科学引文数据库》、《中国科技论文统计源数据库》等国家重点检索机构列为引文期刊,并被英国《科学文摘》(SA)、俄罗斯《文摘杂志》(AJ)、日本《科学技术文献速报》(JST)、美国《剑桥科学文摘:材料信息》(CSA: MI)、波兰《哥白尼索引》(IC)、德国《数学文摘》(Zentralbratt MATH)等多种国外重要检索系统列为来源期刊.我刊主要刊登内容有先进计算、网络与通信、信息与网络安全、数据库与数据挖掘、人工智能、软件过程技术、图形图像处理、智能感知与识别处理、现代服务业信息技术、典型应用等.

我刊是您学习计算机应用理论,借鉴计算机应用技术,参考计算机应用经验的最佳选择.

中国标准连续出版物号: ISSN 1001-9081 / CN 51-1307/TP

国外发行代号: M4616

国内邮发代号: 62-110

定 价: 28 元/册

联 系 人: 雍 平

通信地址: 四川成都 237 信箱(武侯区)《计算机应用》编辑部(610041)

电 话: (028) 85224283

传 真: (028) 85222239-816

电子邮箱: bjb@computerapplications.com.cn

网 址: www.computerapplications.com.cn

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net