EDOLOIS:高效准确的子空间局部离群点发现*

周红福+, 钱卫宁, 魏 藜, 周傲英

(复旦大学 计算机科学与工程系,上海 200433)

EDOLOIS: Efficient Discovery of Local Outliers in Subspaces

ZHOU Hong-Fu+, QIAN Wei-Ning, WEI Li, ZHOU Ao-Ying

(Department of Computer Science and Engineering, Fudan University, Shanghai 200433, China)

+ Corresponding author: Phn: +86-21-65643790 ext 802, E-mail: hfzhou@fudan.edu.cn, http://www.fudan.edu.cn

Received 2003-05-10; Accepted 2004-07-16

Zhou HF, Qian WN, Wei L, Zhou AY. EDOLOIS: Efficient discovery of local outliers in subspsaces. *Journal of Software*, 2004,15(Suppl.):106~113.

Abstract: For many KDD applications, such as data cleaning, detecting criminal activities in E-commerce, etc. finding the outlier can be more meaningful and interesting than finding the common cases. In the paper, we present a novel and efficient subspace local outlier test algorithm: EDOLOIS, so as to avoid the computation-intensive distance computation. The algorithm takes full use of the character of subspace data processing and the initial LOF itself, thus it can not only reduce the computation dramatically, but also gain the precise LOF of all objects in the subspaces. Both formal analysis and comprehensive performance evaluation show that the method is efficient to find all local outliers from high-dimensional categorical datasets in all subspaces.

Key words: data mining; outlier; local outlier; subspace

摘 要: 离群点检测在数据挖掘方面是一项很重要的技术,它是要发现那些行为异常的少量数据,这在数据挖掘的许多领域都有很强的现实意义,如金融欺诈、网络监控等领域.给出了一个高效准确的子空间局部离群点发现的算法(efficient discovery of local outliers in subspaces,简称 EDOLOIS),来避免距离计算的高代价.算法充分利用了原始 LOF 的信息和特点,结合子空间和原空间的关系,从而能够精确且高效地算出子空间局部离群系数,进而甄别出离群点.形式的分析和严格证明都揭示了该算法对在高维种属属性的数据集中发现局部离群点是高效精确的.

关键词: 数据挖掘;离群点;局部离群点;子空间

越来越多的数据存放在数据库中,这使得利用这些信息并从中高效获取知识成为一种当然的需要.长期以来,人们把注意力都放到了那些识别数据集中大量对象有统一模式的方面.事实上,发现数据集中行为异常的少量数据对象,本身就有着很强的应用背景,如网络监控、金融欺诈、数据清洗等领域.

* Supported by the National High-Tech Research and Development Plan of China under Grant No.2002AA413310 (国家高技术研究发展计划(863))

作者简介: 周红福(1977－),男,安徽郎溪人,博士生,主要研究领域为数据挖掘,流数据管理;钱卫宁(1976－),男,博士,讲师,主要研究领域为 P2P 数据管理,流数据挖掘;魏藜(1978－),女,博士生,主要研究领域为数据挖掘;周傲英(1965－),男,博士,教授,博士生导师,主要研究领域为数据挖掘与流数据处理,XML 数据管理,对等计算.

目前,对离群点的研究得到了越来越多的重视.研究者们提出了多种离群点定义并开发了相应的检测算法.对于不同数据区域之间密度差别较大的数据而言,分析和实验表明基于密度的局部离群点定义比较有效;而对于高维数据,由于它们在整个空间中的分布往往比较稀疏,只有在投影子空间中定义离群点才能克服"维度灾难(curse of dimensionality)".现实生活中的数据,往往兼有以上两种特性,因此在寻找离群点时,需要同时考虑数据对象的"局部"和"子空间"的性质.简单地将目前已有的寻找局部离群点和子空间离群点的方法结合起来是不可行的.一方面,局部离群点发现算法要计算空间中对象两两之间的距离,其时间代价为 $O(n^2)$,而且对于大规模数据集,计算距离需要多遍扫描数据集,代价昂贵.另一方面,维数组合爆炸(combinational explosion)使子空间个数以指数级增长,更加剧了这一问题.在仔细分析了局部点定义和高维种属(categorical)属性数据特性的基础上,我们的分析表明:在所有子空间中进行代价高昂的距离计算是不必要的.

本文提出一种针对高维种属属性的数据集,在子空间中进行局部离群点发现的方法 EDOLOIS(efficient discovery of local outliers in subspaces),它有如下特性:能够找到在所有子空间中的所有局部离群点;根据数据在高维的信息高效的计算出数据在低维的离群系数;在子空间中的局部离群点的定义,准确地描述了数据的离群性质,我们的算法主要利用了投影生成的子空间和原空间的关系,同时结合 LOF 本身的定义的特点,从而形式上推出子空间离群系数的计算关系式,进而判断离群点.

本文第 1 节简单介绍了现有的离群点发现方法.第 2 节详细描述了基于密度的局部离群点定义,并举例说明其不足,寻找子空间局部离群点的具体算法在第 3 节中给出.第 4 节是算法分析和结论.最后,第 5 节总结全文,并给出了本文的后续工作.

1 相关工作

到目前为止,离群点还没有一个正式的、为人们普遍接受的定义.Han 的定义[1]揭示了离群点的本质:"经常存在一些数据对象,他们和数据的一般模型不符合;总的来说,他们和数据的其他部分不同伙不一致,我们称其为孤立点("Very often, there exist data objects that do not comply with the general behavior or model of the data. Such data objects, which are grossly different from or inconsistent with the remaining set of data, are called outliers.")".

对离群点发现问题的研究首先是在统计学领域中开展的[2],相关技术可分为基于分布(distance-based)的和基于深度(depth-based)的两大类.基于分布的方法假定数据集满足某一分布,然后根据事先给定的分布来测试离群点,与分布偏差较大的点被认为是离群点.基于深度的方法根据某些统计信息来定义数据的深度,并认为深度较小的数据对象是离群点的可能性比较大.然而,研究表明,这些方法或者需要事先知道数据的分布情况,或者对多/高维数据的处理效率较低.由于数据挖掘的对象往往是复杂的大规模数据集,基于统计的方法很难有效且快速地找到对用户或应用有意义的离群点.

聚类(clustering)算法(如 CLARANS[3],DBSCAN[4],BIRCH[5],CURE[6])也经常考虑离群点对簇(cluster)的影响.但现有的大多数聚类算法的目标是发现包含了大量相似数据对象的簇,并保证聚类质量不受噪声(noise)或离群点影响,因此这些算法并没有给予离群点以足够的重视.一些算法把离群点作为聚类过程的副产品,另一些算法则直接把离群点当作噪声忽略了.

近年来,人们对离群点的价值的认识日益深入.研究者们开展了许多专门针对离群点发现的研究.Knorr 和 Ng 首先提出了基于距离(distance-based)的离群点的概念[7,8].他们认为,如果一个点与数据集中大多数点之间的距离都大于某个阈值,即这个点周围是稀疏的,那么它是一个离群点.基于这个定义,研究者们开发了基于小方格的(cell-based)[7,8]和基于分区的(partition-based)算法[9]用于在 4-维以下的环境中高效地从大规模数据库中发现离群点.

基于距离的离群点定义是全局的.Breunig 等人仔细研究了这种定义[10],得出结论:全局定义不能很好地找到离群点.在此基础上,他们引入了基于密度(density-based)的局部离群点的概念.该定义通过比较一个点与它附近点的密度来考察点的离群程度.他们定义了每个对象的局部离群系数(LOF)以表示离群程度.离群系数越大,对象是离群点的可能性越大.为了快速找到局部离群点,Breunig 等人利用了 OPTICS[11]算法来计算 LOF[10],Jin

等人则利用微簇(micro-cluster)来寻找 n 个 LOF 最大的局部离群点[12].

由于维度灾难对算法性能和有效性的影响,多维/高维空间中的离群点检测显得尤为困难.当维数增多时,数据的分布变得稀疏,这使得多维/高维空间中数据之间的距离尺度及以此为基础的区域密度不再具有直观的意义.聚类研究中,在子空间中分析高维数据性质是一种解决维度灾难的常用方法[13],这些算法的成功证明了数据在不同的子空间中可能具有不同的聚集性质.一些新的研究借鉴这一思想,将高维空间的数据投影到子空间以后再进行离群点检测.

值得注意的是,局部离群点的概念不同于子空间离群点.局部离群点是指一个点相对于数据集的一部分数据(即数据集的一个子集)是特殊的,是对数据集进行横向分割后的一种观察.而子空间离群点是指数据集在某些维的投影上的特殊的点,是对数据集的纵向分割.我们认为:单考虑横向或纵向,都是不够的,会漏掉一些有价值的观察离群点的性质.局部离群点不再具有全局的基于距离的离群点在超空间/子空间中的性质[8].本文致力于解决高效地寻找所有在子空间中的局部离群点的问题.

2 问题描述

本文用 D 表示数据集,o,p,q 表示数据集中的对象,$d(p,q)$表示对象 p,q 之间的距离.相关工作中已提到,全局离群点定义在数据集中存在各种不同密度的簇的情况下是不适用的[10].为了解决这个问题,文献[10]提出了基于密度的局部离群点概念,简述如下.

定义 1. 对象 p 的 k-距离和 k-最近邻

对象 p 的 k-距离 k-dist(p),是对象 p 与数据集 D 中对象 o 之间的距离 $d(p,o)$,如果:

1. 至少有 k 个对象 $o'\in D$ 满足 $d(p,o')\le d(p,o)$ 并且
2. 至多有(k–1)个对象 $o'\in D$ 满足 $d(p,o')<d(p,o)$.

对象 p 的 k-最近邻 $N_k(p)$ 是数据集 D 中与 p 的距离不超过 k-dist(p)的对象集合:

$$N_k(p)=\{q\in D\setminus\{p\}\mid d(p,q)\le k-dist(p)\}.$$

定义 2. 对象 p 的局部密度和局部离群系数

对象 p 的局部密度 $den_k(p)$ 是对象 p 的 k-最近邻中对象的平均 k-dist 的倒数:

$$den_k(p)=1/avg\{k-dist(q)\mid q\in N_k(p)\}.$$

对象 p 的局部离群系数 $LOF_k(p)$ 是对象 p 的 k-最近邻中对象的平均密度与对象 p 的密度的比值:

$$LOF_k(p)=\frac{avg\{den_k(q)\mid q\in N_k(p)\}}{den_k(p)}.$$

从定义易得,对象的 k-最近邻的平均 k-距离越小,密度越大,反之亦然.对象的局部离群系数实际上是它附近对象的密度与它自身密度的比较,离群系数越大,它为离群点的可能性就越大.

这一定义很好地抓住了离群的本质,能够有效地找到相对部分数据行为异常的局部离群点.但是这一定义是在整个空间中给出的,并没有考察数据投影到子空间后的情况.我们看下面的例子.

例 1:图 1 给出的是一个简单的包含 12 个对象的二维数据集,假定每个对象到与它最近的对象之间的距离都是相等的.根据局部离群点的定义,在二维空间中不存在离群点.数据沿 y-轴投影后,$C1$ 与 $C2$ 处因为是多个对象的叠加,密度远高于 o_1' 和 o_2',因此 o_1' 和 o_2' 是一维子空间中的离群点.

这个例子说明,对象在整个空间中不是离群点,不代表它在子空间中也不是离群点.为了找出人们感兴趣的所有离群点,我们应在每个子空间中寻找局部离群点.显然,子空间个数是组合爆炸的,例如一个 20 维的数据集,有 20 个 19 维的子空间,20*19 个 18 维的子空间,....,一般的,n 维空间有 C_n^l 个 l 维子空间.要在这么多的子空间中计算每个对象的 k-距离、k-最近邻、密度及离群系数,计算代价非常大.本文研究的问题就是,如何高效地发现在所有子空间中的所有局部离群点.

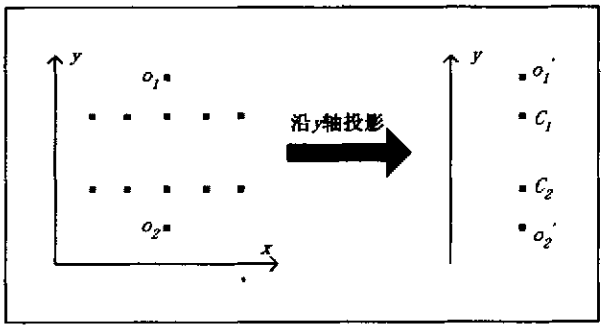

图 1 一个简单的二维数据集

3 子空间离群点精确检测算法

以上分析表明,在子空间中寻找离群点是必要的,但其计算复杂度非常高.本节首先给出一些定义和结论,根据 LOF 的特点并结合子空间和原空间固有的关系,理论推出 EDOLOIS 算法在降一维、两维后子空间离群系数的情形,然后给出针对降 k 维后子空间高校局部离群点检测 EDOLOIS 算法.该算法对子空间离群系数的计算不需要重复进行复杂的距离计算,而是利用原空间已算的结果,从而大大减少计算的代价.

数据集 D,设其有 n 维种属属性,其中的数据对象 o 与 p 之间的距离用广义海明距离来定义,即

$$d^n(o,p)=\left\|\{i|o_i\neq p_i, i=1,2,\ldots,n\}\right\|.$$

定义 3. $\text{Level}_k^n(p,x)$和 $\text{Inner}_k^n(p,x)$,其中 $x\in Z$.$\text{Level}_k^n(p,x)$表示在 n 维空间中,到对象 p 的距离为 k-dist(p)+x 的点集;$\text{Inner}_k^n(p,x)$表示在 n 维空间中,到对象 p 的距离小于 k-dist(p)+x 的点集.这里在 $\text{Level}_k^n(p,x)$和 $\text{Inner}_k^n(p,x)$中 $x(x\in Z)$代表层数,我们称 $\text{Level}_k^n(p,x)$和 $\text{Inner}_k^n(p,x)$分别为膜 x 层和里 x 层.

例如:集合 $N_k^n(p)$ 可分为膜 0 层 $\text{Level}_k^n(p,0)$和里 0 层 $\text{Inner}_k^n(p,0)$;$\text{Level}_k^n(p,0)$是在 n 维空间中数据集 D 中至对象 p 的距离为 k-dist(p)+0 的集合;$\text{Inner}_k^n(p,0)=N_k^n(p)-\text{Level}_k^n(p,0)$.

针对膜层和里层及 LOF 的定义,我们有下面的性质:

性质 1. $\left\|\text{Level}_k^n(p,0)\right\|\geq 1$, $\left\|\text{Inner}_k^n(p,0)\right\|\leq k-1$; $\left\|\text{Level}_k^n(p,0)\right\|+\left\|\text{Inner}_k^n(p,0)\right\|\geq k$.

引理 1. 已知在 n 维空间中对象 o,p 的距离为 $d^n(o,p)$,p 的 k-距离为 k-dist(p),n-l 维子空间中的相应值为 $d^{n-l}(o,p)$ 和 $k-dist^{n-l}(p)$,则有以下不等式成立:

$$d^n(o,p)-l\leq d^{n-l}(o,p)\leq d^n(o,p) \tag{1}$$

$$k-dist^n(p)-l\leq k-dist^{n-l}(p)\leq k-dist^n(p) \tag{2}$$

3.1 降一、二维后子空间的离群系数的精确推导

我们的目的是高效的而准确地推算出低一、二维子空间的 $k-dist^{n-1,2}(p)$ 和 $N_k^{n-1,2}(p)$,从而得到 $den_k^{n-1,2}(p)$ 和 $LOF_k^{n-1,2}(p)$.

3.1.1 降一维后子空间离群系数的精确推导

为了表述方便,我们假设是投影到第 i 维,这样来推算 p 点降一维的离群系数;本文中,点和对象无特别说明,含义相同,都是指数据集中的元素.

引理 1. 降一维的的子空间中,k-dist $^{n-1}(p)$=k-dist(p)或者 k-dist(p)−1(这里作一点申明,如果距离等式为负时,如无其他说明,视为 0;下同)

引理 2. 降一维的的子空间中, $N_k^{n-1}(p)\subseteq N_k^n(p)+\text{Level}_k^n(p,1)$,本文中,集合间的"+"操作同∪.

下面给出如何精确计算降一维子空间的 k-dist $^{n-1}(p)$和 $N_k^{n-1}(p)$.

计算过程:

Step 1. Level$_k^{n-1}(p,1)$←Level$_k^n(p,0)$,Level$_k^{n-1}(p,0)$←Level$_k^n(p,-1)$,Inner$_k^{n-1}(p,0)$←Inner$_k^n(p,-1)$;

Step 2. 在数据集 D 中,在第 i 维上,比较集合 Level$_k^n(p,0)$中的所有对象和 p 是否相异,将该维上相异的对象放入 Level$_k^{n-1}(p,0)$中,同时从 Level$_k^{n-1}(p,1)$除去该对象.

Step 3. 如果 $\|\text{Inner}_k^{n-1}(p,0)\|+\|\text{Level}_k^{n-1}(p,0)\|\geq k$,那么 $k\text{-}disnt^{n-1}(p)=k\text{-}dist(p)-1$, $N_k^{n-1}(p)\leftarrow \text{Inner}_k^{n-1}(p,0)+\text{Level}_k^{n-1}(p,0)$,转到 Step5;否则

Step 4. $k\text{-}dist^{n-1}(p)=k\text{-}dist(p)$,比较集合 Level$_k^n(p,1)$中的对象和 p 是否相异,将相异的对象放入 Level$_k^{n-1}(p,1)$; $N_k^{n-1}(p)\leftarrow$ Inner$_k^{n-1}(p,1)$+Level$_k^{n-1}(p,1)$,Inner$_k^{n-1}(p,1)$=Level$_k^{n-1}(p,0)$+Inner$_k^{n-1}(p,0)$;

Step 5. 由计算的 $k\text{-}dist^{n-1}(p)$和 $N_k^{n-1}(p)$ 就可计算 $den_k^{n-1}(p)$ 和 $LOF_k^{n-1}(p)$.

这里的 $N_k^{n-1}(p)$,表示在 n 维空间种 p 点沿第 i 维投影后在子空间中的对应结点的 k 最近邻在原空间上的象.因此,计算 $den_k^{n-1}(p)$ 是要把 $N_k^{n-1}(p)$ 里的对象与原空间中 p 点的已知的距离求和然后减去该集合中在 i 维与对象 p 在 i 维相异的数目,就是原空间 p 点投影后在子空间与 k 最近邻的距离和.下文同.

3.1.2 降二维后子空间离群系数的精确推导

为了表述方便,我们假设是投影到第 i,j 维,这样来推算 p 点降二维的离群系数.同降一维相似,有关 $k-dist^{n-2}(p)$ 和 $N_k^{n-2}(p)$ 我们有类似的结论:

引理 3. 降二维的的子空间中,k-dist$^{n-2}(p)=k$-dist(p) or k-dist(p)−1 or k-dist(p)−2.

引理 4. 降二维的的子空间中, $N_k^{n-2}(p)\subseteq N_k^n(p)$+Level$_k^n(p,1)$+Level$_k^n(p,2)$.

下面给出如何精确计算降二维子空间的 $k\text{-}dist^{n-2}(p)$和 $N_k^{n-2}(p)$.

计算过程:

Step 1. Level$_k^{n-2}(p,t)$←Level$_k^n(p,t-2)$,Inner$_k^{n-2}(p,0)$←Inner$_k^n(p,-2)$,(t=0,1,2);

Step 2. 在数据集 D 中,在第 i,j 维上,

Step 2.1 比较集合 Level$_k^n(p,-1)$中的所有对象和 p 是否相异,以 q 代 Level$_k^n(p,-1)$中一对象,如果相异,那么 Level$_k^{n-2}(p,1)$减去 q;

Step 2.1.1 仅有一维相异 Level$_k^{n-2}(p,0)$←q;

Step 2.1.2 有二维相异 Inner$_k^{n-2}(p,0)$←q;

Step 2.2 比较集合 Level$_k^n(p,0)$中的所有对象和 p 是否相异,以 q 代 Level$_k^n(p,0)$中一对象,如果相异,那么 Level$_k^{n-2}(p,2)$减去 q;

Step 2.2.1 仅有一维相异 Level$_k^{n-2}(p,1)$←q;

Step 2.2.2 有二维相异 Level$_k^{n-2}(p,0)$←q;

Step 3. 如果 $\|\text{Leve}_k^{n-2}(p,0)+\text{Inner}_k^{n-2}(p,0)\|\geq k$,那么 $k\text{-}dist^{n-2}(p)=k\text{-}dist$(p)−2, $N_k^{n-2}(p)\leftarrow$ Leve$_k^{n-2}(p,0)$+Inner$_k^{n-2}(p,0)$转到 Step 7;否则,

Step 4. 在数据集 D 中,在第 i,j 维上,

Step 4.1 比较集合 Level$_k^n(p,1)$中的所有对象和 p 是否相异,以 q 代 Level$_k^n(p,1)$中一对象,如果相异,那么

Step 4.1.1 仅有一维相异 Level$_k^{n-2}(p,2)$←q;

Step 4.1.2 有二维相异 Level$_k^{n-2}(p,1)$←q;

Step 5. 如果 $\|\text{Leve}_k^{n-2}(p,1)+\text{Inner}_k^{n-2}(p,1)\|\geq k$,其中 Inner$_k^{n-2}(p,1)$ =Level$_k^{n-2}(p,0)$+ Inner$_k^{n-2}(p,0)$;那么 $k\text{-}dist^{n-2}(p)=k\text{-}dist(p)-1$, $N_k^{n-2}(p)\leftarrow$ Leve$_k^{n-2}(p,1)$+Inner$_k^{n-2}(p,1)$,转到 Step 7;否则,

Step 6. 在数据集 D 中,在第 i,j 维上,

Step 6.1 比较集合 Level$_k^n(p,2)$中的所有对象和 p 是否相异,以 q 代 Level$_k^n(p,2)$中一对象,如果二维

相异,那么 $\text{Level}_k^{n-2}(p,2)\leftarrow q$;$k\text{-}dist^{n-2}(p)=k\text{-}dist(p)$, $N_k^{n-2}(p)\leftarrow \text{Leve}_k^{n-2}(p,2)+\text{Inner}_k^{n-2}(p,2)$,

其中 $\text{Inner}_k^{n-2}(p,2)=\text{Level}_k^{n-2}(p,1)+\text{Inner}_k^{n-2}(p,1)$;

Step 7. 由计算的 $k\text{-}dist^{n-2}(p)$和 $N_k^{n-2}(p)$ 就可计算 $den_k^{n-1}(p)$ 和 $LOF_k^{n-1}(p)$.

3.2 降m维后子空间离群点精确检测算法

引理 5. 降 m 维的子空间中,$k\text{-dist}^{n-m}(p)\in\{k\text{-dist}(p)-i|(i=0,1,\ldots,m)\}$;

引理 6. 降 m 维的子空间中,$N_k^{n-m}(p)\subseteq N\text{-cand}_k^{n-m}(p)$,其中

$$N\text{-}cand_k^{n-m}(p)\leftarrow N_k^n(p)+\sum_{i=1}^{m}\text{level}_k^n(p,i).$$

我们把降一、二维后精确检测所生成子空间离群点的算法扩展到降 m 维的情形,初始,令 $\text{Level}_k^{n-m}(p,i)=\text{Level}_k^n(p,i-m)$,$\text{Inner}_k^{n-m}(p,i)=\text{Inner}_k^n(p,i-m)$,具体算法如下:

PROCEDURE EDOLOIS //发现降指定 m 维后子空间的离群点

输入:数据集 D,维数 n,最小点数 k,离群系数的阈值θ,指定 $m(m<n)$维投影;

输出:指定 m 维投影后 n-m 维子空间中的局部离群点

1. $d=n$; //初始化
2. 在 d-维空间中计算每个对象 p 的 $k-dist^d$, N_k^d, den_k^d, LOF_k^d;
3. 若 $LOF_k^d(p)\geq\theta$,p 为 d-维空间中的离群点; //检测离群点
4. IF (d==1) EXIT;
5. ELSE BEGIN
6. $\text{Level}_k^{n-m}(p,t)\leftarrow\text{Level}_k^n(p,t-m)(t=0,1,\ldots,m)$;$\text{Inner}_k^{n-m}(p,0)\leftarrow\text{Inner}_k^n(p,-m)$;$a=k\text{-dsit}(p)-m$// 初始化
7. for (i=0;i<=m;i++)
8. {
9. if (i>0)可递推算出 $\text{Level}_k^{n-m}(p,i)=\text{Level}_k^{n-m}(p,i-1)+\text{Inner}_k^{n-m}(p,i-1)$;// if i==0 $\text{Level}_k^{n-m}(p,0)$和 $\text{Inner}_k^{n-m}(p,0)$都已被初始化.

If (i>0)$\text{Inner}_k^{n-m}(p,i)=\text{Level}_k^{n-m}(p,i-1)+\text{Inner}_k^{n-m}(p,i-1)$;//同样可递归算出

10. for (j=i;j<=i+m;j++)
11. {
12. $\text{Level}_k^n(p,j-m)$中的对象 q 与对象 p 在本次循环所对应的那被指定的 k 维中,if (它们在对应维上相异,设个数为 x 个)//比较 $\text{Level}_k^n(p,j-m)$中的所有对象.
13. {
14. if (j<=m) $\text{Level}_k^{n-m}(p,j)$去掉 q;
15. if $x>j-i$; add 对象 q to $\text{Inner}_k^{n-m}(p,i)$;//if 距离<$a+i$, then 就是修改里 i 层,否则修改膜层;add operation 不应使集合有重复对象
16. else if ($j-m$<=x) add 对象 q to $\text{Level}_k^{n-m}(p,j-x)$;
17. }
18. }
19. if ($\left\|\text{Inner}_k^{n-m}(p,i)\cup\text{Level}_k^{n-m}(p,i)\right\|\geq k$)
20. {
21. $k\text{-dist}^{n-m}(p)=a+i$;
22. $N_k^{n-m}(p)\leftarrow\text{Inner}_k^{n-m}(p,i)+\text{Level}_k^{n-m}(p,i)$;

23. break;
24. }
25. }
26. 由 $k-dist^{n-m}$ 和 N_k^{n-m} 在 $n-m$ 维空间中精确计算出每个对象 p 的 $den_k^{n-m}(p)$ 和 $LOF_k^{n-m}(p)$
27. if $LOF_k^{n-m}(p)>\theta$ 对象 p 就是离群点,输出 p.
28. ENDBEGIN

ENDPROCEDURE

4 结论和算法分析

首先,我们先说明一点,在算法 EDOLOIS 中,为了表述方便而使用了指定 m 维,其实完全可以不指定,而在 Step5 后套一层 C_n^m 投影的循环,这样就得到了所有的 $n-m$ 维子空间所有对象离群系数.

该算法在计算子空间的离群点时,所需要的计算量和所降的维数 m 密切相关,所以 m 不适合取得超过 $n/2$,程序中忽略这一点,是为了说明该算法的通用,m 取 1,2 等较小的数,EDOLOIS 有很好的性能,这由下面的分析可知.

设数据集中有 t 个对象,则计算其中对象两两之间距离的代价为 $O(t^2)$.得到了空间距离后,计算一个对象的 $k-dist^d$, N_k^d, den_k^d, LOF_k^d 的代价很低,只需查找满足某些条件的对象或对相应的值做求和、平均等运算,可认为计算复杂度为 $O(1)$.设候选集 $N\text{-cand}_k^{n-m}(p)$ 中的对象为 s,那么寻找候选集的时间为 $O(s)$.根据 EDOLOIS 算法,最坏情况的代价为 $O(m)*(O(s)+O(m)O(s))$,即 $O(sm^2)$,整个子空间就为 $O(tsm^2)$.这就是计算降 m 维后一个子空间所有对象离群系数的代价,显然,选取合适的 m 后,比在子空间先算距离,再求 LOF 的代价 $O(t^2)$要好的多.另外,因我们假设原空间的 $k-dist^d$, N_k^d, den_k^d, LOF_k^d 都是知道的,所以不把算原空间的计算量算进计算子空间的代价,特此作一说明.

5 总结和展望

近年来,数据挖掘界围绕离群点的发现问题开展了很多研究工作.其中,针对高维种属属性的数据集,基于密度的局部离群点和针对高维数据的子空间离群点的定义由于其对各自特性数据的离群性质的准确描述被广泛接受.多数实际数据兼具多样化和高维两种特性,因此综合考虑局部和子空间两方面的性质是十分重要而有意义的.但是,高维数据将带来的"维度灾难"和"组合爆炸"问题导致直接结合局部离群点发现和子空间离群点发现在计算上是不可行的.本文针对高维种属属性的数据集,提出了一种有效的在子空间中进行局部离群点检测的 EDOLOIS 算法.算法利用子空间和原空间固有的数学关系,结合 LOF 定义的特点,利用原空间的原始数据来高效的计算出子空间的局部离群系数.理论分析表明,EDOLOIS 算法能够很好的利用原空间已经计算出的数据,来推算出子空间的局部离群系数,从而较好的克服了"维度灾难"和"组合爆炸"问题对离群点发现问题的影响;理论上,EDOLOIS 算法能够以小的多的时间复杂度发现所有子空间中的局部离群点.

当前,我们的研究工作针对种属属性的高维数据,利用了广义海明距离.对于任意距离尺度,由于在子空间中,距离的收缩不再具有有界的性质,所以当前方法不完全适用.我们正在进行的工作包括两个方面,一方面,是对于距离收缩无界情况下的局部离群系数的准确估计;另一方面,是对于原空间的不同子空间的离群系数之间是否存在能够可以减少计算量的关系,包括利用索引、汇总信息进行更快速的计算.

致谢 在此,我们向对本文的工作给予支持和建议的同行,尤其是复旦大学计算机科学与工程系周傲英教授领导的讨论班上的同学和老师表示感谢.

References:

[1] Han JW, Kamber M. Data Ming: Concepts and Techniques. Beijing: Higher Education Press, 2001. 381.

[2] Barnett V, Lewis T. Outliers in Statistical Data. 3rd ed., John Wiley & Sons, 1994.

[3] Ng RT, Han JW. Efficient and effective clustering methods for spatial data mining. In: Proc. of the 20th Int'l Conf. on Very Large Data Bases. Santiago: Morgan Kaufmann Publishers, 1994. 144~155.

[4] Ester M, Kriegel HP, Sander J, Xu X. A Density-Based algorithm for discovering clusters in large spatial databases. In: Proc. of the 1996 Int'l Conf. Knowledge Discovery and Data Mining. Portland: 1996. 226~231.

[5] Zhang T, Ramakrishnan R, Livny M. BIRCH: An efficient data clustering method for very large databases. In: Proc. of the 1996 ACM-SIGMOD Int'l Conf. Management of Data. Montreal, 1996. 103~114.

[6] Guha S, Rastogi R, Shim K. Cure: An efficient clustering algorithm for large databases. In: Proc. of the 1998 ACM-SIGMOD Int'l Conf. Management of Data (SIGMOD'98). Seattle, 1998. 73~84.

[7] Knorr EM, Ng RT. Algorithms for mining distance-based outliers in large datasets. In: Proc. of the 24th Int'l Conf. on Very Large Data Bases. New York: Morgan Kaufmann Publishers, 1998. 392~403.

[8] Knorr EM, Ng RT. Finding intensional knowledge of distance-based outliers. In: Proc. of the 25th Int'l Conf. on Very Large Data Bases. Edinburgh: Morgan Kaufmann Publishers, 1999. 211~222.

[9] Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for mining outliers from large data sets. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. ACM Press, 2000. 427~438.

[10] Breunig MM, Kriegel HP, Ng RT, Sander J. LOF: Identifying density-based local outliers. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. Dallas: ACM Press, 2000. 93~104.

[11] Ankerst M, Breunig MM, Kriegel HP, Sander J. OPTICS: Ordering points to identify the clustering structure. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. Philadelphia: ACM Press, 1999. 49~60.

[12] Jin W, Tung AKH, Han JW. Mining top-n local outliers in large databases. In: Proc. of the ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining. 2001. 293~298.

[13] Aggarwal CC, Procopius CM, Wolf JL, Yu PS, Park JS. Fast algorithms for projected clustering. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. ACM Press, 1999. 61~72.

[14] Aggarwal CC, Yu P. Outlier detection for high dimensional data. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. Santa Barbara: ACM Press, 2001. 37~47.