离群点挖掘方法综述*)

薛安荣　姚　林　鞠时光　陈伟鹤　马汉达
（江苏大学计算机科学与通信工程学院　镇江 212013）

摘　要　离群点挖掘可揭示稀有事件和现象、发现有趣的模式，有着广阔的应用前景，因此引起广泛关注。首先介绍离群点的定义、引起离群的原因和离群点挖掘算法的分类，对基于距离和基于密度的离群点挖掘算法进行了比较详细的讨论，指出了其优缺点和发展方向，重点对当前研究的热点——高维大数据量的挖掘、空间数据挖掘、时序离群点挖掘和离群点挖掘技术的应用进行了讨论，指出了进一步研究方向。

关键词　离群点挖掘，局部离群点，子空间，剪枝，空间离群点，高维数据，数据流

Survey of Outlier Mining

XUE An-rong　YAO Lin　JU Shi-guang　CHEN Wei-he　MA Han-da
(School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang 212013, China)

Abstract　The identification of outliers can lead to the discovery of truly unexpected knowledge in areas such as electronic commerce, credit card fraud, and even the analysis of performance statistics of professional athletes. This survey provided a comprehensive overview of existing outlier mining techniques and summarized their features to help users choose, studied and improve algorithms for outlier mining. Studied the outlier mining techniques on high-dimensional data, spatial data and sequential data, pointed out the advantages and disadvantages, and put forward their research direction about outlier mining in future work.

Keywords　Outlier mining, Local outlier, Subspace, Pruning, Spatial outlier, High dimensional data, Sequential data

1　引言

离群点检测(outlier detection)是数据挖掘的基本任务之一[1-6]，故称为离群点挖掘，其目的是消除噪音或发现潜在的、有意义的知识。对离群点挖掘的研究经历了几次盛衰交替，近10年再次成为信息科学中一个活跃的分支，在数据库、数据挖掘、机器学习和统计学等领域受到广泛关注，有着广阔的应用前景，如欺诈检测、入侵检测、故障检测、生态系统失调、公共卫生中的异常疾病的爆发、公共安全中的突发事件的发生、异常自然气候的发现等[1-18]。

离群点有多种别名，如孤立点、异常点、新颖点、偏离点、例外点、噪音、异常物等[1-18]，这里通称为离群点。

引起离群的原因主要包括：①数据来源于异类，如欺诈、入侵、疾病爆发、不寻常的实验结果等。这类离群通常都是相对有趣的，并且是离群检测的关注点。②数据变量固有变化引起，是自然发生的，反映了数据集的数据分布特征，如气候变化、顾客的新的购买模式、基因突变等，这类离群是有趣的。③数据测量和收集误差，主要是由于人为错误、测量设备故障或存在噪音。由于这类离群不提供有趣的信息，只会降低数据和其后数据分析的质量，因此目标是消除这类离群。

早期对离群点研究的主要目的是消除离群点，然而由于“一个人的噪音是另一个人的输入信号”，不加区分而简单地剔除，可能会丢失有趣的重要信息。现在对离群点的研究主要是作为有意义的输入信号，对其进行有效挖掘，以便进一步分析。

离群点挖掘通常可以看作3个子问题：①什么样的数据是异常，即离群点的定义；②有效挖掘离群点的方法；③离群点的意义，即离群挖掘结果的合理解释。

本文目的是对已有离群点挖掘的研究成果进行综合分析，指出潜在应用及进一步研究方向。余下内容组织如下：第2节讨论离群点的分类；第3节和第4节分别讨论基于距离和基于密度的离群点定义、挖掘算法及其局限；第5节指出离群点挖掘的研究热点及发展趋势；最后给出结论。

2　离群点挖掘方法的分类

离群点的挖掘方法很多，可分为5类：基于分布的、基于深度的、基于聚类的、基于距离的和基于密度的[1]。

2.1　基于分布的离群点

离群点检测最早出现在统计领域。基于分布的方法是假设给定的数据集符合某种概率分布模型(例如正态分布)或利用给定的数据集自动构造其概率分布模型，然后根据分布模型采用不一致性检验来确定离群点[7,8]。

到目前为止，还没有一个广为接受的离群点的正式定义，但Hawkins的定义抓住了概念的精髓：“一个离群点是一个观察点，它偏离其它观察点如此之大，以至引起怀疑是由不同机制生成的”[8]。依据该定义可给出基于正态分布的离群点定义。

定义1(基于正态分布的离群点)　设O是关于平均值为

*)基金项目：国家自然科学基金(60603041)，江苏省高校自然科学基金(05KJB520017)。**薛安荣**　CCF会员，博士生，副教授，主要研究方向为数据库与数据挖掘；**姚　林**　硕士生，主要研究方向为数据挖掘；**鞠时光**　CCF高级会员，博士，教授，博士生导师，主要研究方向为数据库与信息安全；**陈伟鹤**　博士，副教授，主要研究方向为数据库与信息安全；**马汉达**　硕士，高级工程师，主要研究方向为网络信息系统。

μ 和标准差为 σ 的正态分布的数据对象集，若 $o\in O$，$|\frac{o-\mu}{\sigma}|\geqslant 3$，则 o 为离群点。

该定义是基于分布的一个具有代表性的离群点定义，偏离平均值 μ 超过 3σ 的数据点就是离群点。文献[7]针对不同的数据分布提出了 100 多种离群点检测方法。检测方法的选择依赖于：1)数据的分布；2)分布参数；3)预期的离群点数目；4)离群数据类型。

基于分布离群点挖掘的主要优点是：1)有坚实的概率统计理论支撑；2)根据概率统计模型，可有效揭示所发现的离群点的含义；3)在模型构造后，不要求基于模型的数据，完全可以只存储描述模型的最少量的信息。其主要缺点是：1)基于分布的绝大多数方法是针对单个属性的，而许多数据挖掘问题要求在多维空间中发现离群点，这个限制使得它们不适合多维数据集；2)基于分布的方法是假设数据符合某种分布规律，因而不适合分布未知的情形。

2.2 基于深度的离群点

由于大多数数据点并非符合某种数据分布，为了改进这种情况，在计算统计中已经发展了一些方法，其中最好的是基于深度的方法[19,20]。基于深度的方法是给每个数据对象分配一个深度值，将数据对象按分配的深度值映射到二维空间的相应层上，处在浅层上的数据对象比深层上的更有可能是离群点[19,20]。基于深度的方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。实际上，现有的基于深度的方法仅对于 $k\leqslant 3$ 其性能可接受[2,3]。

2.3 基于聚类的离群点

基于聚类的算法是先将数据集分成若干簇，不属于任何簇的数据点就是离群点，比较典型的算法有 DBSCAN[21]，CLARANS[22]，CHAMELEON[23]，BIRCH[24]，STING[25]，WaverCluster[26]和 CLIQUE[27]。

定义 2(基于聚类的离群点)　如果一个对象不属于任何簇，则该对象是基于聚类的离群点。

基于聚类离群点挖掘的主要优点是：1)由于对聚类的研究成果比对离群点的研究成果更多，而且有些聚类技术(如 k 均值)的时间和空间复杂度是线性或接近线性的，因而基于这种算法的离群点检测技术可能是可行和高度有效的；2)簇的定义通常是离群点的补，因此可同时发现簇和离群点。其主要缺点是：1)聚类算法的主要目标是发现簇，而不是发现离群点，因此对离群点的挖掘效率较低；2)在聚类过程中，为了避免离群点对聚类的影响，不同算法采用了适合特定数据类型的方法，因此算法的针对性很强，必须小心地选择聚类算法；3)基于聚类的离群点挖掘算法依赖于所有簇的个数和数据中离群点的存在性。

2.4 基于距离的离群点

为了改进上述挖掘方法的缺陷，Knorr 和 Ng[4-6]引入了基于距离的离群点概念和挖掘方法，有效处理了五维以上的大数据集的离群点挖掘问题。但也存在时间复杂度高，需要 $O(\delta N^2)$ 的时间复杂度，其中 δ 为维度，N 为总的数据点个数；对于高维数据，难以解决稀疏问题；挖掘结果对参数的选择非常敏感；由于使用全局阈值，未考虑局部密度的变化，因此只能挖掘全局离群点，不能挖掘局部离群点。

2.5 基于密度的离群点

基于密度的离群点的定义是在距离的基础上建立起来的，将点之间的距离和给定范围内点的个数这两个参数结合起来得到“密度”的概念。一个点的离群程度与它周围的点有关，这体现了“局部”的概念，即局部离群点的概念[10,11]。基于密度的离群点检测给出了对象离群程度的定量度量，对不同密度区域中的数据也能够很好地处理，解决了局部离群点的离群程度的度量和挖掘问题。但仍具有 $O(\delta N^2)$ 时间复杂度，且参数选择比较困难。

基于距离和基于密度的离群点挖掘算法是近 10 年来最具代表的挖掘方法，下面将作进一步分析。

3 基于距离的离群点挖掘

Knorr 和 Ng 给出具有一般意义的基于距离的离群点定义和相应的离群点挖掘方法，该方法不需要明确的数据分布，通过 k 邻居距离来确定是否离群。比较典型的基于距离的离群点定义有以下 3 个[4-6]。

定义 3(DB(pct,D)-Outlier)　如果数据集中至少有 pct 部分对象与对象 o 的距离大于 D，则对象 o 是一个基于距离的关于参数 pct 和 D 的离群点，即 DB(pct,D)-*Outlier*。

从定义可以知道，如果 o 在 D 范围内有不多于 $N(1-pct)$ 个邻居，则 o 是 DB(pct,D)-Outlier。用参数 D 确定对象 o 的邻域，参数 pct 判断对象 o 是否为离群点。

实际上，对于恰当定义的 pct 和 D，一个基于分布的离群点定义同样可以利用 DB(pct,D)来定义，如定义 1 可以用 DB$(0.9988,0.13\sigma)$来表述[4-6]；同时，它克服了基于统计的挖掘方法难以处理多维属性和要求用户预先知道数据集服从哪种统计分布模型的缺点。

基于 DB(pct,D)的挖掘算法有基于索引算法、基于块嵌套循环算法和基于单元的算法[4-6]。基于索引的算法采用多维索引结构(如 R 树或 k-d 树)来查找每个对象 o 在半径 D 范围内的邻居。这个算法在最坏情况下的复杂度为 $O(\delta N^2)$，当维度 δ 增加时，复杂度的增加是线性的。但是，复杂度估算只考虑了搜索时间，而索引结构的构建是非常费时的。

基于块嵌套循环算法和基于索引的算法有相同的计算复杂度，但它避免了索引结构的构建，试图最小化 I/O 的次数，把内存的缓冲区分为两半，将数据集分为若干个逻辑块。通过精心选择逻辑块装入每个缓冲区域的顺序，可改善 I/O 效率。

基于单元格的 DB(pct,D)算法将 δ 维空间划分为边长为 $D/(2\sqrt{\delta})$ 的单元格，并以单元格为单位进行检测。其计算复杂度是 $O(m(2\sqrt{\delta}+1)^{\delta}+N)$，其中 m 是单元个数，因此该算法仅适合于大数据集、低维度的场合。DB(pct,D)对参数 pct,D 比较敏感，而且缺少离群程度的信息，因此难以度量和有效挖掘[1]。

定义 4(top-n Outlier)　如果一个数据集具有 N 个对象，给定对象离群程度的计算公式，计算每个对象的离群得分，离群得分最高的 n 个对象就是所求离群点，即 top-n Outlier。

在定义 3 和定义 4 基础上发展了以下两种定义。

定义 5(top-n D^k-Outlier)　数据集 O 中那些到其第 k 个最近邻居的距离 D^k 最大的 n 个对象就是离群点，即 top-n D^k-Outlier。

若 $D^k(o)$ 表示对象 o 与其第 k 个最近邻居的距离，则处于分布稀疏区域的数据点将具有较大的 D^k 值，而属于聚类中的类内数据点将具有较低的 D^k 值。D^k 离群点挖掘方法基

于各数据点 D^k 的排列，克服了 DB(pct,D)法缺少离群程度信息的不足。同时，D^k 法无须用户指定距离参数 D。

文献[28]给出了一种基于划分的发现算法。首先利用聚类算法划分数据集；然后计算各划分 P 的 D^k 边界(P. lower, P. upper)，使 P 中的每个点 p，满足 $P.lower \leqslant D^k(p) \leqslant P.upper$，并利用此信息确定 P 中是否可能包含离群点；最后仅在可能包含离群点的划分中计算和寻找离群点。由于所要寻找的离群点数目 n 相对较少，该方法可通过排除包含大量数据点的划分而降低计算量。实验显示，该方法关于 N 和 $\delta(\leqslant 10)$的可扩展性均较好。但是，由于 $D^k(p)$并没有包含 p 点所有 k 个最近邻的全部信息，因而它并不能很好地反映其邻域的紧密或稀疏状况。

定义 6(top-n w_k-Outlier)　数据集 O 中那些与其 k 个最近邻居的距离之和 w_k 最大的 n 个对象就是离群点，即 top-n w_k-Outlier。

对于数据点 o，对象 o 与其 k 个最近邻居的距离和称为 o 的权，记为 $w_k(o)$。显然 $w_k(o)$比 $D^k(o)$更精确地度量了 o 的邻域的稀疏程度。

定义 5 和定义 6 利用排序，减少了距离参数 D 的输入，增加了参数 n。输出的离群点的个数受 n 控制，但离群点的顺序不受 n 影响，且易于确定。定义 5 仅考虑了第 k 个邻居的距离而忽略了最近邻居值，定义 6 考虑了所有邻居值，是基于最近邻居密度的计算，虽降低了计算速度，但提高了度量精度。

4　基于密度的离群点挖掘

上述离群点定义是对数据集进行全局观察，离群点挖掘方法均基于各数据点自身的邻域来判别其是否是离群点，其检测标准是全局的、绝对的，因此所挖掘的离群点是全局离群点。但许多实际的数据集结构更复杂，还存在另一种离群，这些离群是相对于它们的局部邻域异常，因而被认为是“局部”离群。图 1 是二维数据集，图中包含两个簇 C_1，C_2 和两个离群点 o_1，o_2，其中 C_1 稠密，C_2 稀疏，o_2 是全局离群点，o_1 是局部离群点。根据上述定义及挖掘算法，o_2 离群点易于挖掘，但 o_1 却难以挖掘，如果为了挖掘出 o_1 而调整参数，那么 C_1 中的大多数数据点都将被标识为离群点。

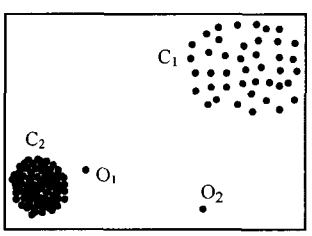

图 1　局部离群示意图

为此，Breunig[10,11]等提出了局部离群点概念和基于密度的离群点定义，通过引入一个专门的度量单位：离群系数(OF：Outlier Factor)，用局部离群系数(LOF：Local Outlier Factor)来表征一个对象的局部离群程度[10,11]。在 LOF 算法[11]中，根据给定的参数最少邻居数 k 和最近邻距离来确定邻域，通过计算对象的 k-距离、可达距离和可达密度，用数据对象邻域的平均可达密度与其自身的可达密度之比表示 LOF，LOF 越大，其离群程度越高。LOF 解决了局部离群程度的度量和挖掘问题，同时摒弃了以往方法中数据对象非此即彼的概念。

图 2 中，由于对象 p，q 的最近邻域密度相同，且对象 q 更靠近 C_1 簇，因此根据 LOF 算法，属于比较稀疏的 C_2 簇的 p 的离群程度高于对象 q，这显然是错误的，为此 Jin 等[29]提出了基于“反向 k 邻域”RNN_k(Reverse k Nearest Neighbors)的局部离群度度量方法 INFLO (INFLuenced Outlierness)，不仅考虑数据点的 k 邻域，还考虑数据点的“反向 k 邻域”对数据离群度的影响，从而避免数据分布复杂情况下 LOF 算法可能出现的错判。采用 INFLO 方法后，在分析 p，q，r 的 k 邻域对象的同时，进一步分析各个对象的 RNN_k，发现 C_2 中对象 s 和 t 的 k 邻域包含点 p，即 s 和 t 属于 $RNN_k(p)$，而 q 的 RNN_k 为空，r 的 RNN_k 仅包含一个点(如图 3[29])。INFLO 方法结合对象 p 的 RNN_k 中对象对 p 的影响，可以得出 p 的离群度小于 q 和 r 的离群度的正确结果。

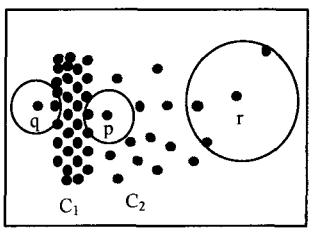

图 2　比较点 p，q，r 的离群程度

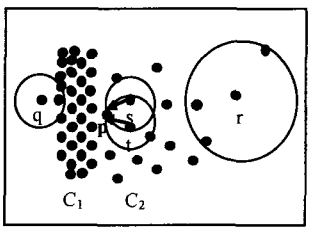

图 3　反向 k 邻居

为了克服 LOF 算法对于序列数据和低密度数据对象不能有效度量的缺陷，Tang 等[30]提出基于连接的离群系数(COF：Connectivity-based Outlier Factor)的方法，其算法是根据给定的参数最少邻居数 k 和数据对象的连接性来确定邻域，计算与其邻域的平均连接距离，用平均连接距离比作为基于连接的离群系数 COF。虽可克服上述局限，但由于 COF 增加了连接路径的建立，因此计算比 LOF 更复杂。

LOF，INFLO 和 COF 等方法解决了局部离群点定义与挖掘问题，但与基于距离的方法相比，其算法更加复杂，效率也较低，时间复杂度为 $O(kN^2)$的计算复杂度，其中 k 为邻居数，N 为数据点总数，难以用于大规模数据集；另外，检测结果对指定的参数 k-邻居的选择很敏感，当 k 值过小时，在离群点彼此接近，形成一个小的离群簇的情况下，会将这个小的离群簇误判为正常数据簇，导致漏检；当 k 值太大时，接近稠密簇的离群点可能会被误判为正常数据点，也会导致漏检。为了得到满意结果，需要反复调整参数 k，而每次调整参数均须重新构造邻域，邻域构造非常费时，具有 $O(kN^2)$的计算复杂度。为了改进算法的效率，降低对参数的敏感性，一些学者提出了避免距离或密度计算的方法、基于划分和剪枝技术的方法、基于属性划分的方法等等，具体方法如下。

Jin[31]基于 LOF 提出了挖掘 top-n 个离群点的思想。算法首先利用微聚类对数据集进行压缩，计算每个微聚类的 LOF 值的上下界，然后在 LOF 值最大的 n 个聚类中查找 LOF 值最大的 n 个对象。Chiu A L 等[32]从另外两个角度对 LOF 进行了推广：其一是将原来的一种邻域推广到了两种邻域，即计算密度的邻域和比较密度的邻域；其二是采用剪除部分非离群对象来减少计算 LOF 的代价。Malik[33]为了降低 LOF 的计算复杂度，提出用局部稀疏系数(LSC：Local Sparsity Coefficient)来表示对象的离群程度。在 LSC 的计算中，用局部稀疏比取代 LOF 中局部可达密度，而用 LSC 取代 LOF，省去了局部可达距离的计算，降低了计算的复杂度。

Papadimitriou[34]引入多粒度偏差因子(MDEF: Multi-granularity DEviation Factor)来度量对象的离群程度。在MDEF算法[34]中,有两个邻域概念,即 r-邻域和 αr-邻域,其中 $r>0, 0<\alpha<1$。$n(p_i, r)$ 和 $n(p_i, \alpha r)$ 分别表示以 p_i 为圆心、r 为半径和 αr 为半径的圆内的对象数目,$\hat{n}(p_i, r, \alpha)$ 表示在 p_i 的 r-邻域内的所有对象 p 的 $n(p, \alpha r)$ 的平均值。MDEF的定义如下:

$$MDEF(p_i, r, \alpha)=\frac{\hat{n}(p_i, r, \alpha)-n(p_i, \alpha r)}{\hat{n}(p_i, r, \alpha)}=1-\frac{n(p_i, \alpha r)}{\hat{n}(p_i, r, \alpha)} \quad (1)$$

MDEF算法的优点是可以根据应用要求设置多级邻域,并用邻域中包含的对象数目替代距离计算,降低了计算复杂度,但 r 和 α 的确定依然比较困难,检测结果和计算复杂度在一定程度上取决于用户的经验。

赵科平等[35]提出了基于相邻关系(NOF: Neighborhood-based Outlier Factor)的离群点挖掘算法。与LOF和COF定义相比,NOF更直观和简单,它仅仅考虑数据对象 p 的 k 邻居的数目 $|kNB(p)|$ 和将 p 作为 k 邻居的数据对象的数目 $|R-kNB(p)|$。简单的NOF定义为

$$NOF(p)=(|kNB(p)|+1)/(|R-kNB(p)|+1) \quad (2)$$

从而避免了直接计算距离或密度。在实验数据集上,NOF算法显示出了比LOF更高的效率。由于本质上仍然是 D^k 值,不是真正的密度,所以难以精确度量点 p 周围的密度。

李存华等[36]采用数据空间网格化方法实现对密集数据主体的过滤,算法对于低维数据具有良好的时间和空间开销,但由于高维数据网格划分不可避免地造成低密度超方格的大量存在,因此在处理高维数据时性能急剧下降。

薛安荣等[1]为了提高离群度度量的精度和挖掘的效率,提出属性划分的方法,属性划分后利用多维索引技术,使计算复杂度由 $O(kN^2)$ 降为 $O(kN\log N)$,通过属性权值的分配减少了冗余属性的干扰,提高了挖掘精度。

5 离群点挖掘研究热点及发展趋势

当前离群点研究主要以距离或密度来计算离群度,研究的重点是高维大数据、空间数据、时序数据和实际应用。

5.1 高维大数据集中离群点的挖掘

随着采集设备性能的提高和数量的增加,采集数据的维数和数量均呈上升趋势,有些数据的维数甚至高达上百维,这对已有离群点挖掘算法是一个挑战。因为,现有的挖掘方法大多是基于数据之间的相似度来挖掘离群点,而在高维情况下,数据十分稀疏,数据点之间的距离及区域密度不再具有直观的意义,因此上述算法对高维离群点的挖掘不再有效或效率比较低。事实上,基于相似的定义,稀疏的高维数据隐含每一个点几乎都可能是很好的离群点。因此,对高维数据而言,发现有意义的离群点也变得十分复杂和不明显。Aggarwal等[37]提出了高维空间中值得思考的几点建议:①处理好高维空间中数据的稀疏问题;②合理地解释离群点产生的原因;③选择合适的度量方法,以解释 d 维子空间中离群点的物理意义;④高维数据的离群点挖掘的计算效率;⑤判断一个点是否为离群点时,要考虑到数据点的局部行为。

为了解决高维离群点挖掘问题已经提出了降低维度[37-39]和重新设计距离函数[40]的办法。降低维度技术主要包括投影变换[37-40]和属性提取[37-39]等方法。投影变换是将数据集从原 δ 维投影到 d 维空间,其中 $d\ll\delta$,并且每个新维是原始维的线性组合,然后在 d 维空间上利用传统的挖掘算法进行挖掘。由于具有 δ 维属性的数据集,可能的维数组合数为 2^δ,对于高维数据,这个数字将是一个天文数字。Aggarwal等[37]提出了用遗传算法寻找最优子空间,优化解决该问题的办法,实验测试了包括对具有279个属性的高维数据集的离群点挖掘,取得良好效果。Angiulli等[38,39]提出HilOut算法,利用Hilbert空间填充曲线(Hilbert space filling curve)将数据集线性化,并基于此线性化的数据集上的前驱关系和后继关系,可快速地找出各点的 k 个近似最近邻,避免了直接求解每对点之间的距离;算法中将全维特征空间多次投影到[0,1]区间,每一次投影都改善了离群点在全维空间中的离群度得分,这样先求近似解,然后从中获取精确解的策略。Yu等[41]利用小波变换(Wavelet Transform)的多分解特性,从原始数据集中消除聚类,从而达到发现离群点的目的。Dutta[42]等使用主成分分析(PCA: Principal Component Analysis)方法获得能代表数据的 δ 维属性的 d 个最正交向量(属性),投影变化后再进行挖掘。

投影变换方法的主要局限是在变换后的子空间上挖掘的结果难以解释,而且为了不丢失信息,需要在不同的子空间上挖掘,这就出现低维子空间的重叠和数据对象重复出现在不同子空间上的问题。

另一种减少维度的方法是特征选取方法。这种方法不用变换,而是从维度中启发式地选取一部分维,删除不相关或冗余的属性(维),目标是找出最小属性集,使得数据类的概率分布尽可能接近使用所有属性得到的原分布。这种方法避免了挖掘结果难以解释问题,并且由于属性数目的减少,使得模式更易于理解。基于启发式方法的技术包括逐步向前选择、逐步向后删除、向前选择和向后删除的结合和决策树归纳[2,3]。许龙飞[43]等利用粗糙集的属性约简技术减少高维空间的维数,并在各个关联规则子空间下对数据集进行基于密度的离群点挖掘,使高维空间下的离群点挖掘更具有实用性。该算法在子空间中使用了LOF的思想进行离群点挖掘,其近邻的判断仍然是基于距离的计算,因而只对数值型数据有效。

上述方法只是在固定子空间上挖掘,而离群点可能在特定的小的子空间内最有意义,这就意味着不同属性在离群点的挖掘中扮演着不同角色[44,45],起着不同作用,使用投影变换或特征选取获得的子空间并非上述特定的小的子空间。因此薛安荣等[1]提出属性划分的方法,将属性空间划分为环境子空间(context subspace)和固有子空间(inherent subspace),环境子空间是指对象行为属性发生的时间、地点和序列位置等属性,固有子空间是指对象的行为属性或特征属性。环境属性决定了对象与其外部的关联,可用这类属性来确定对象的邻域,而固有属性决定了对象的行为特征,可用该属性计算对象的离群度。据此解决了高维对象难以利用多维索引来提高搜索效率的问题,通过属性划分和给不同属性分配不同权值(0~1)等方法减少了与计算比较不相关属性的干扰,提高了挖掘精度。但属性的划分和权值的分配需要一定的领域知识。

总之,现有研究还局限在特定数据类型或特定背景环境,计算还比较复杂。要达到实用,还必须在理论与技术上做进一步研究。

5.2 空间离群点的挖掘

由于GPS、卫星、CT成像等各种宏观与微观传感器的使用,空间数据的数量、大小和复杂性都在飞快地增长,出现"空

间数据爆炸但知识贫乏"的现象。空间数据比关系数据复杂，具有空间自相关性和异质性特点，其属性按性质可分为空间维属性和非空间维属性。空间对象经常受到邻近对象的影响，因此空间离群点挖掘只有充分考虑了对象的邻近点的影响才能获得有用的知识。空间离群点是指那些非空间属性值和邻域中其它空间对象的非空间属性明显不同的空间对象，两个空间对象的差异程度通常用相异度来衡量。由于空间数据自身的特殊性，空间离群点一般是局部不稳定的，这种局部意义上的离群点在全局中不一定仍为离群点。空间离群点挖掘在地理信息系统、遥感图像数据勘测、公众安全与卫生、交通控制、基于地理位置的服务等各种领域有着广泛的应用[13]。

空间离群点的挖掘首先出现在空间统计学中，主要方法可分为图形检测和代数检测两类，如变差云图(variogram cloud)法和 Z-Score 法[9]。但这些方法由于没有考虑空间数据的特点，没有区分空间和非空间属性，其检测效果不佳[13]。Shekhar 等首先提出将空间属性与非空间属性区分开来的二分算法[9,46-48]，并通过对象与其邻域的非空间属性值之差或之比，来消除空间的自相关性，并用该值表示对象与其邻域的偏差。然而该方法未能很好地解决空间的异质性问题，所使用的阈值全局统一，因此挖掘的是全局离群点，不是真正的空间离群点。由于上述算法是针对单维非空间属性的，因此 Lu[49] 等及文俊浩等[50] 提出用 Mahalanobis 距离来解决多维非空间属性的相异度的计算及空间离群点挖掘算法。用 Mahalanobis 距离虽可解决多维属性的相异度度量，但由于使用的阈值仍是全局的，因此挖掘的仍是全局离群点。Chawla 等[51] 同时考虑了空间的自相关性和异质性，用欧氏距离来消除空间对象与其邻域间的自相关性，引入波动参数 β，并用 β 和对象与其邻域的欧拉距离的乘积表示空间局部离群度 SLOM(Spatial Local Outlier Measure)。但由于 β 仅由对称分布状况来决定，在空间邻居较少或波动幅度较小的情况下难以准确表现波动情况，因此出现漏检和误检现象。当 β 不起作用时，退化为基于距离的离群点挖掘算法，所求的是全局离群点，不是真正的空间离群点。Kou 等[52] 在距离计算时考虑了邻居的影响程度，将权重因子加入到距离计算中。薛安荣等[13,53] 提出基于空间约束的离群点挖掘算法，算法中用计算邻域距离的方法解决空间自相关性约束问题，用计算空间局部离群系数 SLOF(Spatial Local Outlier Factor)的方法解决空间异质性约束问题。用对象的邻域距离与邻域中对象的平均邻域距离之比表示空间离群系数，据此挖掘离群点。实验结果表明，在挖掘精度、用户依赖性和计算效率方面取得了比较好的效果。

随着传感器设备技术的发展，数据采集设备的数量越来越多，精度越来越高，因此数据量越来越大，维数越来越高，提高算法的有效性及计算的高效性仍然是空间离群点挖掘算法的发展方向。

5.3 时序离群点的挖掘

时序数据是指按时间顺序取得的一系列观测值。一般地，这些观测值是在等单位间隔时间里采集到的，典型的例子包括某地区的月降雨量、每月的用电量、网络流量等。时序数据中的离群点可能隐含在季节性或其它周期性变化之中，使得离群点挖掘变得更为复杂。但由于其巨大的应用潜力，引起了越来越多研究者的关注[15,16,18,54-62]。传统的时间序列离群模式挖掘一般有两种方法：一种是将时间序列分成等长的子序列，并将子序列映射为 d 维空间中的点，然后采用基于距离的挖掘算法发现离群点。这种方法的一个缺点是序列中的点一般较多，距离的计算和检测的时间消耗是相当可观的。另一种方法是从时间序列中抽取特征，通过计算特征序列间的距离来发现异常，如 AR(自回归)模型及其改进的 ARMA(自回归滑动平均)模型。模型表示法的一个缺点是事先要假定某个模型，而实际上，用户很难确定所要分析的时间序列服从什么模型。Jagadish 等[54] 采用信息论的方法给出了时间序列中离群点的定义框架，并提出了一种在时间序列中挖掘离群点的有效算法。Choy[55] 提出了一种适合大样本、静态时间序列的基于频谱的离群点检测算法 SODA(Spectrum-based Outlier Detection Algorithm)，该算法可用来挖掘定时的、类型确定的离群事件。Ma 等[56] 将一种支持向量机(SVM: Support Vector Machine)方法应用于时序离群点挖掘中，其思想是先将时间序列投影到一个向量空间，然后使用 SVM 进行离群点挖掘。

Dasgupta 等[57] 提出采用人工免疫系统的负选择机制挖掘时序离群点的方法。通过建立自我和异己模型，然后由负选择机制辨别自我和异己。该方法的主要缺点是挖掘结果依赖于自我和异己模型，出现漏检与误检率较高。Shahabi 等[58] 采用基于小波的树型结构 TSA-Tree 表示不同尺度的时序数据，将时序数据中的突变定义为离群点，通过小波系数的局部极大值来发现。该方法由于不能全面准确地反映各种情况，因此出现漏检现象。

Keogh[59]、Bejerano[60] 和 Sun 等[61] 相继提出用概率后缀树(PST: probabilistic suffix tree)存储挖掘的序列结点信息，并用剪枝技术减少处理的数据量，取得了比较好的挖掘效果和挖掘效率。

5.4 离群点挖掘的应用

在自然界、人类社会或数据集中，大部分事件和对象是平凡的或平常的，然而敏锐地捕获到不寻常或不平凡的对象或事件，有着极为重要的意义[3]，如干旱威胁农作物、运动员的超常能力可能致胜、实验结果的异常可能意味着新的现象的出现、对异常现象的进一步研究，可能出现新的理论或新的元素。离群点挖掘的应用前景广阔，下面给出 3 个应用方面，但决不限于此。

(1)欺诈检测。盗窃信用卡的人，其购买行为可能不同于信用卡持有者。信用卡公司根据信用卡消费数据建立信用卡持有者的购买行为模式，异常检测机制及时发现不同寻常的消费行为，通过与持有者的交互可确认是否是欺诈行为，这样可使损失降低到最小。寻求有效的欺诈检测解决方案已经迫切地提上了诸如信用卡公司、银行、保险公司、电信公司、航空公司等商业公司的议事日程上[3,63,64]。类似的方法可用于其它类型的欺诈检测。

(2)入侵检测。入侵不同于系统的正常行为，这个特性允许直接将该问题转化为离群点检测问题，离群点检测技术已经广泛应用在入侵检测中[2,3,16,64,65]。

离群点检测在入侵检测中应用的关键挑战是巨大的数据量、高维的数据属性和实时的在线分析和较低的假警告率要求。入侵检测一般采用半监督和无监督的离群点检测技术。

(3)异常气候的检测。在自然界中，异常气候的出现可能预示着自然灾害的发生，如地震、干旱、洪水、飓风、热浪和火灾。时空离群点的挖掘可解决异常气候发现的问题，预测这些事件的似然度和它们的成因[9,46-49]。

结束语 离群点挖掘技术有着广阔的应用前景，已引起越来越多的关注，但由于离群点的定义还未统一，缺乏通用的测试数据集和对测试结果好坏的通用的衡量标准，制约了离群点挖掘技术的发展。此外，快速有效地发现海量高维数据集中的离群点仍是比较复杂的问题，至今没有通用、有效的方法。

本文通过对常用离群点挖掘方法，特别是基于距离和基于密度的离群点挖掘方法的分析讨论，指出了其优劣和进一步发展方向。并且就目前离群点挖掘研究的热点和难点进行了讨论，指出离群点挖掘未来研究的重点将是高维大数据集、空间数据集、时序数据集的快速有效的挖掘研究，隐私保护的离群点挖掘技术以及实际应用的研究。

参 考 文 献

[1] 薛安荣，鞠时光，何伟华，等. 局部离群点挖掘算法研究. 计算机学报，2007，30(8)：1455-1463

[2] Han Jiawei，Micheline K. Data mining：concepts and techniques. 2nd edition. San Francisco：Morgan Kaufmann Publishers，2006

[3] Tan Pang-Ning，Michael S，Vipin K. Introduction to data mining. New York：Addison-Wesley，2006

[4] Knorr E，Ng R. Algorithms for mining distance-based outliers in large datasets // Proc. of the 24th VLDB Conference. New York，1998：392-403

[5] Knorr E，Ng R. A Unified Approach for Mining Outliers：Properties and Computation // Proc. of Knowledge Discovery and Data Mining(KDD'97). Newport Beach，1997：219-222

[6] Knofr E，Ng R，Tucakov V. Distance-based outliers：algorithms and applications. The VLDB Journal，2002，8(3/4)：237-253

[7] Barnett V，Lewis T. Outliers in Statistical Data. 3 rd edition. New York：John Wiley and Sons，1994

[8] Hawkins D. Identification of outliers. London：Chapman and Hall，1980

[9] Shekhar S，Chawla S. A tour of spatial databases. Upper Saddle River，N. J.：Prentice Hall，2003

[10] Breunig M M，Kriegel H P，Ng R T，et al. OPTICS-OF：identifying local outliers // Proc. of the 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases. Lecture Notes in Computer Science 1704，Prague，1999：262-270

[11] Breunig M，Kriegel H P，Ng R，et al. LOF：Identifying density-based local outliers // Proc. of ACM SIGMOD Conference. Dallas，2000：93-104

[12] 魏藜，宫学庆，钱卫宁，等. 高维空间中的离群点发现. 软件学报，2002，13(2)：280-290

[13] 薛安荣，鞠时光. 基于空间约束的离群点挖掘. 计算机科学，2007，34(6)：207-210

[14] 李翠平，李盛恩，王珊. 一种基于约束的多维数据异常点挖掘方法. 软件学报，2003，14(9)：1571-1577

[15] 郑斌祥，杜秀华，席裕庚. 一种时序数据的离群数据挖掘新算法. 控制与决策，2002，17(3)：324-327

[16] 赵泽茂，何坤金，陈鹏. Web日志文件的异常数据挖掘算法及其应用. 计算机工程，2003，29(17)：195-197

[17] 汪加才，张金城，江效尧. 一种有效的可视化孤立点发现与预测新途径. 计算机科学，2007，34(6)：200-203

[18] 杨宜东，孙志挥，朱玉全，等. 基于动态网格的数据流离群点快速检测算法. 软件学报，2006，17(8)：1796-1803

[19] Ruts I，Rousseeuw P. Computing Depth Contours of Bivariate Point Clouds. Journal of Computational Statistics and Data Analysis，1996，40(23)：153-168

[20] Johnson T，Kwok I，Ng R. Fast Computation of 2-dimensional Depth Contours // Proc. of the 4th KDD. New York，1998：224-228

[21] Ester M，Kriegel H P，Sander J，et al. A density-based algorithm for discovering clusters in large spatial databases with noise // Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining. Portland，1996：226-231

[22] Ng R T，Han J. Efficient and effective clustering methods for spatial data mining // Proc. of the 20th VLDB Conference. Santiago，1994：144-155

[23] George K，Han Eui-Hong(Sam)，Vipin K. CHAMELEON：A Hierarchical Clustering Algorithm Using Dynamic Modeling. IEEE Computer：Special Issue on Data Analysis and Mining，1999，32(8)：68-75

[24] Zhang T，Ramakrishnan R，Linvy M. BIRCH：an efficient eata clustering method for very large databases // Proc. of the ACM SIGMOD International Conference on Management of Data. Montreal，1996：103-114

[25] Wang W，Yang J，Muntz R. STING：a statistical information grid approach to spatial data mining // Proc. of the 23rd VLDB Conference. Athens，1997：186-195

[26] Sheikholeslami G，Chatterjee S，Zhang A. WaveCluster：a multi-resolution clustering approach for very large spatial databases // Proc. of the 24th VLDB Conference. New York，1998：428-439

[27] Agrawal R，Gehrke J，Gunopulos D，et al. Automatic subspace clustering of high dimensional data for data mining applications // Proc. of the ACM SIGMOD International Conference on Management of Data. Seattle，1998：94-105

[28] Ramaswamy S，Rastogi R，Kyuseok S. Efficient algorithms for mining outliers from large data sets // Proc. of the ACM SIGMOD International Conference on Management of Data. Dallas，2000：427-438

[29] Jin Wen，Tung Anthony K H，Han Jiawei，et al. Ranking Outliers Using Symmetric Neighborhood Relationship // Proc. of the PAKDD. 2006：577-593

[30] Tang J，Chen Z，Fu A，et al. Enhancing effectiveness of outlier detections for low-density patterns // Proc. of the 6th PAKDD. Taipei，2002：535-548

[31] Jin Wen，Tung A K，Han Jiawei. Mining Top-n Local Outliers in Large Databases // Proc. of the KDD'01. San Jose，2001：293-298

[32] Lai-mei C A，Wai-chee F A. Enhancements on Local Outlier Detection // Proc. of the 7th International Database Engineering and Applications Symposium. Hong Kong，2003：298-307

[33] Agyemang M. Local Sparsity Coefficient-based Mining of Outliers. Windsor Ontario：University of Windsor，2003

[34] Papadimitirou S，Kitagawa H，Gibbons P B，et al. LOCI：Fast outlier detection using the local correlation integral[A] // Proc. of the 19th International Conference on Data Engineering. Bangalore，2003：315-326

[35] 赵科平，周水庚，关佶红，等. 一种新的离群数据对象发现方法 // 中国人工智能学会第10届全国学术年会论文集. 北京：北京邮电大学出版社，2003

(下转第27页)

[6] Sig2dat specification. http://www.geocities.com/vlaibb/,2002
[7] Cornelli F. Choosing reputable servents in a P2P network[C]// Lassner D, ed. Proc. of the 11th Int'l World Wide Web Conf. Hawaii:ACM Press, 2002:441-449
[8] Kamvar S D, Schlosser M T. EigenRep: Reputation management in P2P networks[C]// Proceedings of the 12th Int'l World Wide Web Conference. Budapest: ACM Press:123-134
[9] Zhang Zhen,et al. A P2P Global Trust Model Based on Recommendation[C]//Proceedings of the Fourth International Conference on Machine Learning and Cybernetics. Guangzhou, August 2005
[10] Stoica I, et al. Chord: A Scalable Peer-to-peer Lookup Service for Internet Applications[C]//SIGCOMM'01. San Diego, California, USA, August 2001
[11] Ratnasamy S. A Scalable Content-Addressable Network[C]// SIGCOMM'01. San Diego, California, USA, August 2001
[12] Joseph D, Kubiatowicz J D. Routing Algorithms for DHTs: Some Open Questions[C]// Electronic Proceedings for the 1st International Workshop on Peer-to-Peer Systems (IPTPS '02) 2002. MIT Faculty Club, Cambridge, MA, USA,2002
[13] 窦文.构造基于推荐的 Peer-to-Peer 环境下的 Trust 模型[J].软件学报,2004,15(4):571-580

(上接第 18 页)

[36] 李存华,孙志挥,陈耿.基于网格上近似的大规模数据集离群点检测算法 GROUT.计算机应用研究, 2003, 20(9):34-136
[37] Aggarwal C C, Yu P. Outlier detection for high dimensional data// Proc. of the ACM SIGMOD International Conference on Management of Data. Santa Barbara, 2001:37-47
[38] Angiulli F, Pizzuti C. Outlier Mining in Large High Dimensional Data Sets. IEEE Trans. Knowledge and Data Eng., 2005, 2 (17):203-215
[39] Angiulli F, Basta S, Pizzuti C. Distance-based detection and prediction of outlier. IEEE Trans. Knowledge and Data Eng., 2006, 2(18): 145-160
[40] Aggarwal C C. Re-designing Distance Functions and Distance-based Applications for High Dimensional Data. SIGMOD Record Date, 2001, 30(1):13-18
[41] Yu Dantong, Gholamhosein S, Zhang Aidong. FindOut: Finding Outliers in Very Large Datasets. Knowledge and Information Systems, 2002,4(4):387-412
[42] Dutta H, Giannella C, Borne K, et al. Distributed top-k outlier detection in astronomy catalogs using the demac system// Proc. of 7th SIAM International Conference on Data Mining. Minneapolis,2007:208-215
[43] 许龙飞,熊君丽.基于粗糙集的高维空间离群点发现算法研究.计算机工程与应用,2004,40(7):58-60
[44] Knorr E M, Ng R T. Finding Intentional Knowledge of Distance-based Outliers// Proc. of the 25th VLDB. Edinburgh,1999:211-222
[45] Chen Zhixiang, Tang Jian, Fu Ada Wai-Chee. Modeling and Efficient Mining of Intentional Knowledge of Outliers// Proc. of the 7th International Database Engineering and Applications Symposium Conference. Hong Kong,2003:44-53
[46] Shekhar S, Lu C-T, Zhang P. A Unified Approach to Spatial Outliers Detection. GeoInformatica, 2003,7(2):139-166
[47] Shekhar S, Lu C-T, Zhang P. Detecting Graph-based Spatial Outliers. International Journal of Intelligent Data Analysis (IDA), 2002, 6(5):451-468
[48] Lu C-T, Chen Dechang, Kou Yufeng. Algorithms for Spatial Outlier Detection// Proc. of 3rd International Conference on Data Mining. Melbourne, 2003: 597-600
[49] Lu C-T, Chen Dechang, Kou Yufeng. Detecting Spatial Outliers with Multiple Attributes// Proc. of the 15th International Conference on Tools with Artificial Intelligence. Sacramento, 2003:122-128
[50] 文俊浩,吴中福,吴红艳. 空间孤立点检测. 计算机科学,2006, 33(5):185-187
[51] Sanjay C, Sun Pei. SLOM: a new measure for local spatial outliers. Knowledge and Information Systems, 2006, 9(4): 412-429
[52] Kou Y, Lu C-T, Chen D. Spatial Weighted Outlier Detection// Proc. of the SIAM Conference on Data Mining. Bethesda, 2006: 613-617
[53] Xue Anrong, Ju Shiguang. Algorithm for Spatial Outlier Detection Based on Outlying Degree// Proc. of the WCICA 2006. Dalian,12(7):6005-6009
[54] Jagadish H V, Koudas N, Muthukrishnan S. Mining deviants in a time series database// Proc. of the 25th VLDB. Edinburgh, 1999:341-350
[55] Choy K. Outlier detection for stationary time series. Journal of Statistical Planning and Inference, 2001, 99 (2):111-127
[56] Ma J, Perkins S. Time-series novelty detection using one-class support vector machines// Proc. of the International Joint Conference on Neural Networks. 2003:168-175
[57] Dasgupta D, Forrest S. Novelty detection in time series data using ideas from immunology// Proc. of the International Conference on Intelligent Systems. 1999:82-87
[58] Shahabi C, Tian X, Zhao W. TSA-tree: a wavelet-based approach to improve the efficiency of multi-level surprise and trend queries// Proc. of the 12th International Conference on Scientific and Statistical Database Management. 2000:55-68
[59] Keogh E, Lonardi S, Chiu B. Finding surprising patterns in a time series database in linear time and space// Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, 2002:550-556
[60] Bejerano G, Yona G. Modeling protein families using probabilistic suffix trees// Proc. of the Third Annual International Conference on Computational Molecular Biology. 1999:15-24
[61] Sun Pei, Chawla S, Arunasalam B. Mining for Outliers in Sequential Databases// Proc. of the Sixth SIAM International Conference on Data Mining. Bethesda, 2006: 94-105
[62] 薛安荣,何伟华. 基于时序离群检测的新的分段方法.计算机工程与设计,2007,28(20):4875-4877
[63] 姚卫新. 智能数据分析中异常数据的集成化管理方法研究.上海:复旦大学,2004
[64] 陆声链. 孤立点挖掘及其内涵知识发现的研究与应用. 南宁:广西大学,2005
[65] Gwadera R, Atallah M J, Szpankowski W. Reliable detection of episodes in event sequences. Knowledge and Information Systems, 2005, 7(4):415-437