第 22 卷 第 1 期 天 文 学 进 展 Vol.22, No.1
2004 年 3 月 PROGRESS IN ASTRONOMY Mar., 2004

文章编号: 1000-8349(2004)01-0001-09

离群数据的探测

张彦霞, 赵永恒

(中国科学院 国家天文台, 北京 100012)

摘要: 综述了离群数据 (outliers) 探测是数据挖掘和知识发现的一项重要任务及其在天文学中兴起的必然性。简要介绍了离群数据的定义、特点、产生原因及影响, 着重阐述了探测一维离群数据和多维离群数据的方法, 并且与一些聚类算法作了对比。每一种算法各有优劣, 天文学家应根据天文数据的特点, 探讨出适合天文数据特点的离群数据探测方法, 以发现一些不同寻常的、稀有的、甚至新类型的天体和天文现象。

关 键 词: 数据处理; 离群数据; 综述; 数据挖掘

中图分类号: P1; N37 **文献标识码**: A

1 引 言

随着天文数据的飞速增长, 亟待强而有效的分析方法来充分挖掘隐藏在数据中的信息。数据库中的知识发现 (Knowledge Discovery in Databases, KDD) 就是从数据中提取隐含在其中、人们事先不知道、但又是潜在有用的信息和知识的过程。有关这方面知识可参考文献 [1] 。按照发现模式的种类不同, KDD 任务也各不相同, 通常分为 4 类 [2,3] : (1) 依赖性探测 (如相关规则) ; (2) 种类的证认 (如分类和聚类) ; (3) 种类的描述 (如概念的推广) ; (4) 例外或离群数据 (outliers) 探测。前 3 种任务主要是根据数据中大部分数据的特征来分类或作模型证认。大部分的数据挖掘研究属于这 3 类。例如: 聚类的目的就是发现一组种类或类别来描述整个数据结构; 分类的目的是找到一个函数将每一个数据点映射到几种给定的类别中。与这 3 种任务不同, 另一种重要的 KDD 任务的研究对象是数据中那些偏离大部分数据分布的数据点, 即例外和离群数据。例外和离群数据在 KDD 领域一直未引起足够重视, 仅仅将其作为噪音副产品而被忽略或抛弃。一些机器学习和数据挖掘的算法尽管考虑了离群数据的存在, 但也只是在某种程度上容忍它们的存在。然而, 在某些情况下, 甲的噪音恰是乙的信号。事实上, 生活中不乏这样的例子。例如: 电子商务中发现的罪犯活动、录像监视、药学研究、专业运动员的成绩评估和气象预测等方面, 发现的稀有事件比通常情况更有趣、更有用。天文学也不例外, 发现稀有的、未知种类的天体和天文现象是天文学家尤其关心和关注的课题。

收稿日期: 2003-07-30 ; **修回日期**: 2003-09-18

基金项目: 国家自然科学基金资助项目 (10273011)

2 离群数据

2.1 离群数据的定义

目前还没有一个一般的、统一的、大众普遍接受的离群数据的定义，下面给出几种定义：

(1) Hawkins-Outlier [2,4]

离群数据是指那些观测值远远偏离其它观测值，以至于怀疑它是由其它机制产生的数据。

(2) DB $(pct, dmin)$-Outlier [4]

在数据集 D 中，给定到 p 的距离 $dmin$ ，若至少有百分之 pct 的数据到 p 的距离大于 $dmin$ ，则称数据 p 为 DB $(pct, dmin)$-Outlier 。

(3) Local Outlier [4~6]

离群数据以它邻域的物体为参照，看其离群的可能性，即计算出每个数据点的离群因子，挑选出离群因子最大的几个作为 Local Outlier (局部离群数据) 。

2.2 离群数据的产生原因及其影响

从定义可知，离群数据通常与类别紧密相连：离群数据是那些偏离数据主要分布的数据，换句话说，它们远离或不属于某一类。离群数据具有如下特点：

(1) 数据的预测值是异常的；

(2) 数据的平均值是异常的；

(3) 数据对参数的影响无法事先估计。

如果数据具备这 3 个特点中的一个，通常就是离群数据，研究者需要对其引起重视。一般，产生离群数据的原因有：

(1) 数据项本身的错误，如仪器、天气或人为因素造成观测值的错误；

(2) 样本的非均匀性。一些样本的数目远远小于其它样本，导致这些小样本常被视为离群数据；

(3) 对未知的数据结构作出不正确的分布假设，从而导致原本不是离群数据的数据成为离群数据；

(4) 在误差分布的两翼上，极值出现的频率远大于在预期的正常分布上出现的频率，因而这些极值常被当作是离群数据；

(5) 稀有事件或现象。

离群数据的存在，改变了所在数据组的平均值、方差和回归参量，增大了均方差，使估计或预测出现偏差，从而导致错误的结论。离群数据之所以引起人们兴趣的原因还有两个 [7] ：一是，在数据分析之前，需要平衡各类样本数目。若某类样本数目偏多，而其它样本数目偏少，很可能会将数目少的样本当作离群数据剔除。所以应去掉样本数较多的样本中的离群数据以使各类样本均衡，这类离群数据勿需进一步研究；二是，某些离群数据会给我们新的视角，需要对其认真审视和研究。比如在天文学中一些稀有的、新类型天体的发现，将导致新理论的发展和完善，高红移类星体、褐矮星的发现即属此列。在大的数据组中通过系统地寻找参数空间中的离群数据可以发现稀有的未知类型天体， SDSS (Sloan Digital Sky Survey) 和 DPOSS (Digitized Palomar Sky Survey) 研究组在寻找高红移类星体时

就是利用这种方法发现了一些新类型天体的。 1998 年 Djorgovski 等人[8] 在 DPOSS 巡天中利用颜色参数空间发现了高红移类星体和 II 型类星体。颜色参数空间中，正常恒星分布呈香蕉形，并形成一个温度序列，而类星体不同于正常恒星，它们远离恒星区，如图 1 所示[9] 。根据它们的吸收线和发射线特点，可以区分高红移类星体和 II 型类星体。这两类类星体相当稀少，其面密度每平方度小于 10^{-2} ，低于可靠的恒星与星系分类的界限。这样，为了统计性地探测一些有意义的样本，就需要大天区巡天及合适的选择方法。类似地，这一方法也可应用于其它波段的低角分辨率巡天。例如：利用 IRAS (Infrared Astronomical Satellite) 数据区分恒星和星系，用射电指数把类星体从射电星系中辨别出来，用 X 射线的硬度比挑选 AGN 等；在可见光和近红外波段找到各个红移的类星体的完备样本[10,11] ；选出特殊谱类型的恒星用以探测银河系结构[12] 。如果星系的形态可以参数化，则同样可以将星系按形态区分开[13] 。在未探索的参数空间中，系统地寻找离群数据可以发现另外一些特殊天体，其中一些结果便是新天文现象的最初证据。因此在统计分析、机器学习、模式识别、数据挖掘之前，都需要对数据预处理，挑出离群数据，按其产生的原因进行取舍。若是误差或噪音，则剔除；若是稀有或特殊事件，则需要详细研究。

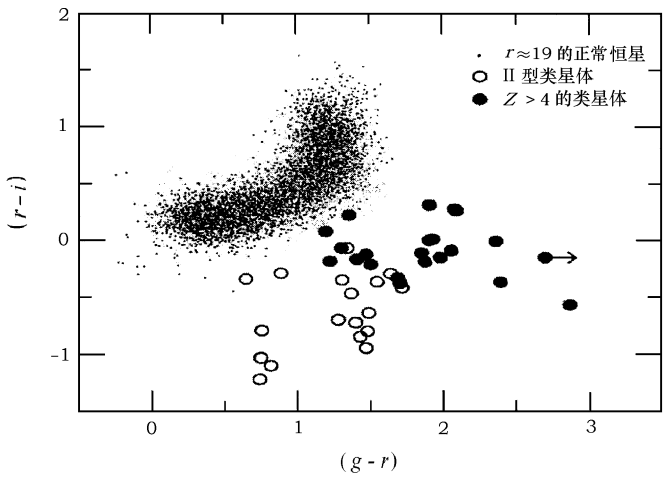


图 1　DPOSS 巡天得到的三色空间中天体的分布[9]

2.3　离群数据的探测方法

按变量来分，离群数据分为单变量离群数据和多变量离群数据。最简单而普遍的离群数据研究集中于单变量离群数据的确认[7] 。在单变量情况下，极值显然是离群数据。当数据分布对称时，左右两边的尾部 (tails) 端点很可能是离群数据。对应两边的点分别称为下离群数据 (lower outliers) 和上离群数据 (upper outliers) ；当数据分布不对称时，分布较宽一侧的尾部极值可能是离群数据。相比之下，多变量的数据分布不存在尾部，因而多变量离群数据的探测比较困难。当然多变量离群数据有时可能恰是单变量离群数据，但这种情况毕竟为数极少；通常一些在一维数据空间中正常分布的数据点在多维数据空间中将成为多变量离群数据。

2.3.1　一维离群数据的寻找方法

以一组数据为例，介绍与离群数据相关的几个定义及几个评判数据是否为离群数据的方

法。现共有 70 个数据，以升序排列如下：

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

(1) 平均值 (Mean)

对某样本，其平均值 $\overline{x}=\frac{\sum_{i=1}^{n}x_i}{n}$，式中 x_i $(i=1,2,\cdots,n)$ 为观测值，n 代表包含的数据点数。

本例的平均值 $\overline{x}=\frac{\sum_{i=1}^{n}x_i}{n}=\frac{34356}{70}=490.8$。

(2) 百分点 (Percentile)

一个数组的第 p 个百分点是指在该数组中至少有百分之 p 的数据小于或等于此值，即百分之 $(100-p)$ 的数据大于或等于此值。假如数据按升序排列，则第 p 个百分点的位置 $i=\frac{p}{100}\times n$。当 i 不是整数时，对其取整，第 p 个百分点即是处于第 i 个位置的数据；当 i 是整数时，第 p 个百分点即为第 i 个和第 $i+1$ 个数据点的平均值。

本例的第 90 个百分点的位置 $i=\frac{90}{100}\times 70=63$，因为 i 是整数，取其值为第 63 和 64 个数据的平均值：$\frac{580+590}{2}=585$。

(3) 中值 (Median)

在一个有序样本中，处于最中间的值即中值 $\overleftrightarrow{x}$。当离群数据存在时，中值对于确定数据中心是一个十分有用的量。

本例的中值是第 50 个百分点，$i=\frac{p}{100}\times n=\frac{50}{100}\times 70=35.5$，$\overleftrightarrow{x}=\frac{475+475}{2}=475$。显然本例的平均值大于中值，说明平均值左边的数据偏多。

(4) 众数 (Mode)

样本中出现频率最高的数据为众数。

本例中 450 出现的次数最多 (7 次)，所以它为众数。

(5) 四分点 (Quartiles)

四分点是具体的百分点，如第 1 个四分点 $(Q1)$ 是第 25 百分点，第 2 个四分点 $(Q2)$ 是第 50 百分点即中值，第 3 个四分点 $(Q3)$ 是第 75 百分点。四分点范围 (interquartile range, IQR) 为 $IQR=Q3-Q1$。其中，内下限：$Q1-1.5\times IQR$；内上限：$Q3+1.5\times IQR$；外下限：$Q1-3.0\times IQR$；外上限：$Q3+3.0\times IQR$。

弱离群数据 (mild outliers) 是小于内下限而大于外下限或大于内上限而小于外上限的数据，强离群数据 (extreme outliers) 则为小于外下限或大于外上限的数据。

在本例中内下限和内上限分别为：$Q1-1.5\times IQR=450-1.5\times 75=337.5$，$Q3+1.5\times IQR=525+1.5\times 75=637.5$。所以本例中不存在离群数据。

(6) 值域 (Range)

值域为最大值与最小值的差值，是一种最简单的弥散检测方法，从其定义很容易看出它极易受到最大值和最小值的影响。

(7) 方差 (Variance)

方差为每一个数值与平均值差值的平方和的平均值，是最重要的衡量变化的尺度。

(8) 标准偏差 (Standard Deviation)

标准偏差是方差的正的平方根。对整个样本，标准偏差为 σ；对某类样本，标准偏差则为 s。

(9) 方差系数 (Coefficient of Variation)

方差系数表示相对于平均值，标准偏差有多大。对于确定变量随不同平均值和标准偏差的变化，方差系数是十分有用的。

本例中，方差 $s^2=\dfrac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}=2996.16$，标准偏差 $s=\sqrt{s^2}=\sqrt{2996.16}=54.74$，方差系数 $\dfrac{s}{\overline{x}}\times 100=\dfrac{54.74}{490.80}\times 100=11.15$。

探测一维离群数据的几种数值方法如下：

(1) z-Scores 法

z-Scores 常称作标准值，它代表观测值相对平均值的偏离与标准偏差的比值，即 $z_i=\dfrac{x_i-\overline{x}}{s}$。

本例中最小值的 z-Scores 为 $z=\dfrac{425-490.80}{54.74}=-1.20$。

离群数据通常是数据集中那些非常大或非常小的数据，z-Score 大于 3 或小于 -3 的数据即被认为是离群数据。在本例中，最大和最小的 z-Score 值分别为 2.27 和 -1.20，因此本例不存在离群数据。

(2) Chebyshev's 定理

由 Chebyshev's 定理可知，概率 $P(|x-\mu|<k\sigma)\geq 1-\dfrac{1}{k^2}$ 成立，其中 μ 为样本的平均值，样本中至少有 $\left(1-\dfrac{1}{k^2}\right)$ 的数据在其平均值的 $k\sigma$ 范围内，k 为大于 1 的整数 [14]。

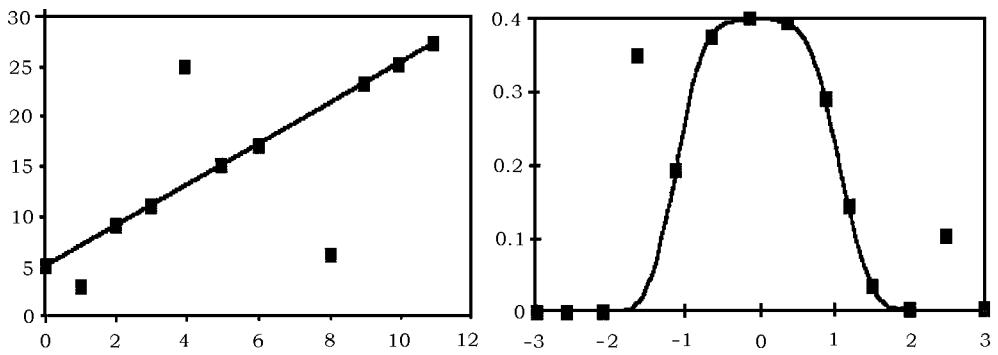
例如：当 $k=2$ 时，至少有 75% 的数据在平均值的 2σ 范围内；当 $k=3$ 时，至少有 89% 的数据在平均值的 3σ 范围内；当 $k=4$ 时，至少有 94% 的数据在平均值的 4σ 范围内。

(3) 经验规则 (Empirical Rule)

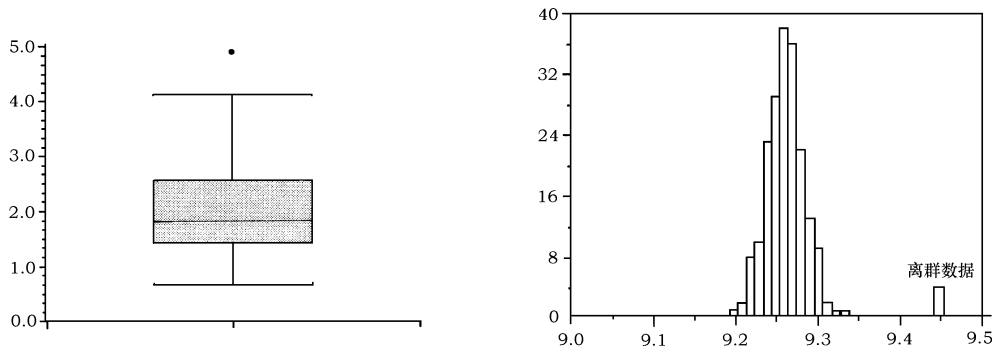
在具有正态分布的数据组中，大约 68% 的数据在平均值的 1σ 范围内，95% 的数据在平均值的 2σ 范围内，99% 的数据在平均值的 3σ 范围内。

为形象直观地发现离群数据，散点图 (scatter plot)、框图 (box plot)、直方图 (histogram) 不失为一种好方法，如图 2 所示。在处理一维离群数据时，Barnett 和 Lewis [15] 针对不同的数据分布 (正态、泊松、指数和二项式分布) 提出了约 100 种不一致检验或离群数据检验的方法。至于具体选择哪一种检验依赖于：1) 数据的分布；2) 分布参数是否已知；3) 预期的离群数据数目；4) 离群数据的种类。但是，所有的检验均存在两个严重的问题。首先，这些检

验针对的数据是单变量的，因而不适合多变量情形；其次，这些检验是以数据分布为基础的，因而不适合分布未知的情形。实际上，在每一次检验前并不知道数据分布属于哪一种类型，例如是正态分布还是伽玛分布，故必须尝试各种检验，以期找到合适的数据分布。



(a) 散点图



(b) 框图　　(c) 直方图

图 2　散点图、框图和直方图

2.3.2　多维离群数据的寻找方法

证认一维离群数据的方法建立在数据的排序上，而对于多维数据，不存在无争议的排序标准。Barnett 和 Lewis [15,16] 提出了一种与众不同的子排序 (sub-ordering) 方法，这一方法在离群数据的研究中被普遍应用。建立简化的子排序方法需两步 [7] 。首先，以距离标量为基础转换每一个多维观测值 x_i 为标量 r_i ，形成一组标量 $R = \{r_i\}(i = 1, \cdots, N)$ 。 R 按实际的多维数据的顺序排列，故离群数据是那些具有较大 R 值的多维观测值。

推广的距离公式如下 [7,15,16] ：

$$r_i^2 = (x_i - x_0)' \Gamma^{-1} (x_i - x_0) \ , \tag{1}$$

式中， x_0 代表数组的平均矢量， Γ^{-1} 表示权重矩阵与数据的弥散成反比。这些参数的不同选择将导致不同的距离度量。例如， Γ 为单位阵时， (1) 式代表 x_i 到数据中心 x_0 的欧几里得距离。

若选用马氏距离 (Mahalanobis distance) [17,18]，只需将 (1) 式中的 Γ 改为协方差矩阵 $\sum$ 即可。通常，中心参量取分布的期望值 μ。然而 μ 常常未知，故一般用样本的平均矢量 $\boldsymbol{m}$ 和协方差矩阵 $\boldsymbol{S}$ 来估计：

$$r_i^2 = (x_i - m)' \boldsymbol{S}^{-1} (x_i - m) \ . \tag{2}$$

马氏距离合并了属性间的依赖关系。这一点在多维离群数据探测中尤为重要，因为其目的就是要探测异常值的合并情形。许多距离度量包括欧几里得距离仅利用中心位置 (location) 信息，因此不适合这种任务。马氏距离的另一个优点是把每一个变量标准化为零均值和单位方差，这样单位方差不会对距离产生影响。

伽玛概率统计分布图在利用推广的距离公式探测离群数据时比较有用。这些图画出了按顺序排列的简约的单变量 r_i 随伽玛分布的四分点变化。如果多维观测值服从正常分布 (如正态分布、泊松分布)，那么简约量 r_i 近似伽玛分布，数据点应围绕一条直线聚类，那些明显偏离线性关系的数据即为离群数据。在不能确定数据分布是否服从正态分布时，最好利用框图。因为伽玛概率统计分布图需要专家来评估反常数据点是否确实为离群数据，而框图则提供了确定离群数据存在的客观标准。尽管离群数据的不一致检验有其坚实的统计理论基础，但必须与它的每一条假设符合。在实际应用中，若一些假设不符合时，利用众所周知的、广泛应用的非正式方法 (如框图) 将是比较合适的选择。当然，将来的工作还是应体现出规范的统计检验。需注意利用马氏距离探测多维离群数据时有两点限制：首先，数据服从定量的正态分布；其次，缺值在计算距离之前应处理。也许某种不同寻常的距离函数可以解决这个问题 [7]。

计算统计学领域发展了许多种适合多维离群数据探测的方法。归纳起来，可分为 5 类 [5,6]。(1) 基于分布的 (distribution-based)，离群数据是那些偏离正常分布 (如正态分布、泊松分布) 的数据；(2) 基于深度的 (depth-based)，该探测方法依赖于计算 k 维凸球的不同层，离群数据是那些处于球外层的数据；(3) 基于距离的 (distance-based)，包括 DB (pct，$dmin$)-Outlier 方法和统一方法 (unified approach for mining outliers)；(4) 基于聚类的 (clustering-based)，如 CLARANS、DBSCAN 和 BIRCH 方法；(5) 基于密度的 (density-based)，如 Breunig 等人 [5] 提出的发现局部离群数据的方法 OPTICS-OF、Jin 等人 [6] 提出的挖掘最离群的 n 个局部离群数据 (mining top-n local outliers) 算法。还有一种是基于偏差的 (deviation-based)，Aggarwal 和 Yu 提出的遗传算法 [19,20]，以及其它方法如贝叶斯方法 [21]、以分形为基础的小波方法 [22]、模糊集理论 [23]、并行算法 [24] 和所有的非监督聚类方法，如自组织映射 (Self-Organization Map，SOM)、主

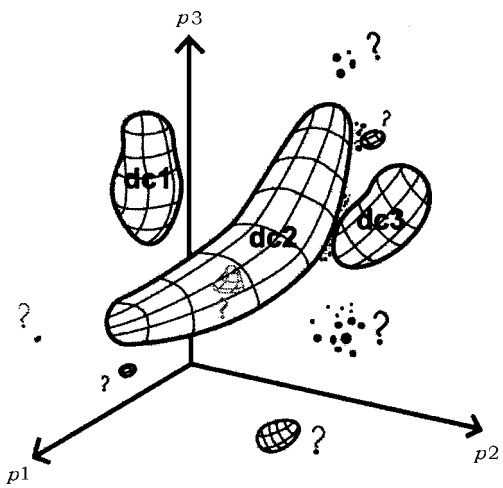


图 3　聚类分析事例

其中 dc1、dc2、dc3 为 3 大主要类型天体，带“？”的数据为一些稀有或未知类型天体、离群数据或主要类别中的空穴.

分量分析 (Principal Component Analysis， PCA) 等。图 3 给出了一个聚类分析的例子 [25] 。该例中数据分布在三维 $p1$ 、$p2$ 、$p3$ 参数空间中，大部分数据属于 dc1 、dc2 、dc3 3 大类。主要类型天体的聚类可用于统计研究，如恒星作为星系结构的探针、星系作为宇宙大尺度结构的探针、挖掘异常天体。对于那些少量的、统计上又很重要的、明显聚为一类的数据，甚至孤立的离群数据，很可能就是稀有的或未知类型的天体。另外主要类别中的空穴也有可能对应一些有趣的物理现象，抑或数据本身有问题。

聚类算法主要是将具有相似属性的事物归为一类,它并不是为探测离群数据而设计的 [3] ，像 CLARANS 、DBSCAN 和 BIRCH 算法是专为数据挖掘设计的聚类算法。在 CLARANS 算法中，如果剔除某一数据项能提高聚类因子，那么该数据项将被当作噪声去掉。同样在 BIRCH 算法中，如果某数据相对较偏离距离它最近的聚类中心，那么也会被当作噪声处理。这两个算法中，离群数据的定义是通过聚类间接定义的。聚类算法的发展是为优化聚类，而非探测离群数据，离群数据仅是聚类分析的副产品。在统计学中，聚类算法通常分为两类：分割 (partitioning) 算法和分层 (hierarchical) 算法。这两种情况下，每一数据项至少属于一类。机器学习的所有聚类算法也是如此。区别于 CLARANS 和 BIRCH 算法， DBSCAN 算法提供了较直接的离群数据确认方法，它通过在 ε 邻域内物体的数目和所考虑的数据的可扩展性 (reachability) 和连通性 (connectivity) ，将数据分为核区、边界和离群，同时设定 ε 足够小以获得较强的聚类。DBSCAN 算法主要也是想产生最大数目的数据分类，而非标出哪些数据是离群数据。因此，在探测离群数据时不能简单地将聚类算法拿来用，毕竟它们的着重点在聚类而非离群数据。真正充分挖掘离群数据仍需要发展与之匹配的挖掘算法。

3 结 论

随着天文仪器和观测手段的进一步提高，天文数据将以 TB 甚至 PB 量级计量，数据维数由几维上升到几十维、几百维，从而对算法的要求日趋严格，对天文学家也提出了新的挑战。显然，天文学家仅掌握本领域的知识是远不够的，需要与统计学家、计算机学家、数据挖掘学家合作，以应对形势发展的需求。目前，大部分探测离群数据的算法是基于低维空间的，适合高维空间的算法还有待进一步探讨和研究。每一算法都有其适用范围，作为天文学家应融合数据挖掘、统计学、机器学习和模式识别等学科的优点，探讨出一些切实有效的适合天文数据特点的挖掘算法，以便挖掘出隐藏在数据中鲜为人知的、稀奇的、新类型的天体和天文现象，推动天文学甚至其它学科的发展和完善。相信不久的将来，国际虚拟天文台 (International Virtual Observatory, IVO) 的创建和运行，将使天文学家可以轻松自如地进行在线数据挖掘。

参考文献:

[1] 张彦霞，赵永恒，崔辰州. 天文学进展， 2002, 20(4): 312
[2] Knorr E M, Ng R T, Tucakov V. Very Large Data Base Journal, 2000, 8(3-4): 237
[3] Knorr E M, Ng R T. http://citeseer.nj.nec.com/232267.html, 1997
[4] Breunig M M, Kriegel H P, Ng R T *et al.* http://www.dbs.informatik.uni-muenchen.de/Publikationen/Papers/LOF.pdf, 2000

[5] Breunig M M, Kriegel H P, Ng R T *et al.* http://www.dbs.informatik.uni-muenchen.de/Publikationen/Papers/PKDD99-Outlier.pdf, 1999
[6] Jin Wen, Tung A K H, Han Jiawei. http://www-faculty.cs.uiuc.edu/~hanj/pdf/kdd01.pdf, 2001
[7] Laurikkala J, Juhola M, Kentala E. In: Lavrac N, Miksch S, Kavsek B eds. The Fifth Workshop on Intelligent Data Analysis in Medicine and Pharmacology, Berlin, http://ai.ijs.si/Branax/idamap-2000-AcceptedPapers/Lauri kkala.pdf, 2000
[8] Djorgovski S G, Gal R R, Odewahn S C *et al.* In: Colombi S, Mellier Y, Raban B eds. Wide Field Surveys in Cosmology, Paris: Editions Frontieres, 1998: 89
[9] Djorgovski S G, Mahabal A A, Brunner R J *et al.* In: Brunner R J, Djorgovski S G, Szalay A S eds. Virtual Observatories of the Future, San Francisco: Astronomical Society of the Pacific, 2001: 52
[10] Wolf C, Meisenheimer K, Röser H J *et al.* A&A, 1999, 343: 399
[11] Warren S, Hewitt P, Foltz C. MNRAS, 2000, 312: 827
[12] Yanny B, Newberg H J, Kent S *et al.* ApJ, 2000, 540: 825
[13] Odewahn S C, Windhorst R, Driver S *et al.* ApJ, 1996, 472: L13
[14] http://www.georgetown.edu/faculty/pag9/econstat/notes/notes3.pdf
[15] Barnett V, Lewis T. Outliers in Statistical Data, New York: John Wiley, 1994
[16] Barnett V. J. R. Statis. Soc. A, 1976, 139: 318
[17] Jain A K, Dubes R C. Algorithms for Clustering Data, New Jersey: Prentice Hall, 1988
[18] Boberg J. Academic Dissertation, Finland: Turku Centre for Computer Science, 1999
[19] http://citeseer.nj.nec.com/482998.html
[20] Crawford K D, Wainwright R L, Vasicek D J. Proc. of the 1995 ACM/SIGAPP Symp. on Applied Computing, Nashville: ACM Press, 1995: 351
[21] Varbanov A. http://www.stat.umn.edu/PAPERS/tech-reports/tr614.ps, 1996
[22] Struzik Z R, Siebes A P J M. http://ftp.cwi.nl/CWIreports/INS/INS-R0008.pdf, 2000
[23] Last M, Kandel A. Proc. of International Conference on Intelligent Technologies, 2001, 2: 292
[24] Hung E, Cheung D W. http://citeseer.nj.nec.com/hung99parallel.html, 2002
[25] http://www.astro.caltech.edu/~george/vo/

The Outlier Detection

ZHANG Yan-xia, ZHAO Yong-heng

(*National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China*)

Abstract: The outlier detection is an important task of data mining and knowledge discovery in database. Its necessities in astronomy are reviewed with the definitions, characteristics, forming reasons and causing effects of outliers being simply introduced. The methodologies to detect univariate outliers and multivariate outliers are summarized, and compared with other clustering algorithms. Since each kind of approach has its own pros and cons, astronomers should reasonably choose methodologies to detect outliers according to the characteristics of astronomical data, so that some unusual, rare or even new astronomical objects and phenomena could be found.

Key words: data analysis; outliers; review; data mining