第 43 卷 第 6A 期
2016 年 6 月

计 算 机 科 学
Computer Science

Vol. 43 No. 6A
June 2016

基于相对密度的孤立点和边界点识别算法

李光兴
(成都农业科技职业学院 成都 611130)

摘 要 根据孤立点是数据集合中与大多数数据的属性不一致的数据,边界点是位于不同密度数据区域边缘的数据对象,提出了基于相对密度的孤立点和边界点识别算法(OBRD)。该算法判断一个数据点是否为边界点或孤立点的方法是:将以该数据点为中心、r 为半径的邻域按维平分为 2 个半邻域,由这些半邻域与原邻域的相对密度确定该数据点的孤立度和边界度,再结合阈值作出判断。实验结果表明,该算法能精准有效地对多密度数据集的孤立点和聚类边界点进行识别。

关键词 邻域,密度,孤立度,孤立点,边界度,边界点
中图法分类号 TP311 **文献标识码** A

Recognition Algorithm of Outlier and Boundary Points Based on Relative Density

LI Guang-xing
(Chengdu Agricultural College, Chengdu 611130, China)

Abstract According to the fact that outlier points are the data that are inconsistent with most of data in a data set, and that boundary points are located on the edge of data area with different densities, an algorithm based on relative density was proposed to determine the outlier and boundary points. Through dividing the neighborhood area, which is centered by this point with a radius of r, into two semi-neighborhood areas, and determining this data point's isolation level and boundary level based on the relative density of these semi-neighborhood areas with the original neighborhood area, a final judgment whether a data point is boundary or outlier point can be made according to the threshold value. Experimental results indicate that this algorithm can effectively and accurately identify the outlier and boundary points from multi-density data sets.

Keywords Neighborhood, Density, Isolation level, Outlier, Boundary level, Boundary points

1 引言

孤立点是指属性值明显不同于其邻近对象,偏离了大多数数据行为或数据模型的异常数据。孤立点中可能蕴含着极为重要的信息,是数据分析的主要对象[1,2],具有较高的理论和实际应用价值。

目前常用的孤立点识别方法包括基于统计分布、基于距离、基于密度和基于网格等方法。基于统计的方法[3,4]是创建一个概率分布模型,根据孤立点出现是一个低概率事件的思想,按数据对象拟合模型的情况来评估该数据对象是否为孤立点。这种方法中模型的建立较困难,需要预先估计数据分布状况。基于距离的孤立点识别[5-7]是通过距离来衡量一个对象与其他对象的偏离程度,当一个对象与其他对象的距离越远,该对象成为孤立点的可能性越大。基于距离的孤立点识别方法无需了解数据集的分布模型,适用于可以计算对象间距离的任意维度的数据集。不利之处在于不同密度的数据集的点与点间的平均距离是不同,若采用全局距离参数,难以处理多密度数据集。基于密度的孤立点算法[8,9]是通过分析一个对象一定范围的密度,按具有高密度邻域的数据对象不是孤立点,具有低密度邻域的数据对象可能就是孤立点的思想确定孤立点。但其存在算法复杂、参数选择困难等问题。基于网格的孤立点算法[10,11]是将存储数据对象的空间划分为等边长的单元格,然后按单元格的坐标与数据对象之间的关系将数据对象映射到单元格中,再以单元格为单位来识别孤立点,基于单元的孤立点识别算法虽然避免了对整个数据空间内每个数据点的判断,降低了算法的时间复杂度,但由于得到的孤立点是以单元格为单位,因此比较粗糙,不够准确。

边界点是指位于不同密度数据区域边缘的数据对象,这些对象的全体就是边界。边界反映了数据分布结构信息和属性差异特征,同时也提供了一种数据分类或聚类的模式[12,13]。BORDER 算法[14]是基于密度的边界点识别算法。若一个对象 p 在某个对象 o 的 k 近邻中,则称对象 o 是对象 p 的反向 k 近邻。利用边界点的反向 k 近邻个数一般小于处于聚类内部对象的反向 k 近邻的个数思想来识别边界点。其数据集的所有对象按它的反向 k 近邻数从小到大排列顺序,并把前 t 个对象作为边界点。因为孤立点和聚类内部的一些点的反向 k 近邻数可能少于边界点的反向 k 近邻数,所以该算法不能精准地从多密度和带孤立点的数据集中提取聚类的边界点。另外,参数 t 的多寡取舍比较困难。IBORA 算法[15]通过计算空间向量的夹角来检测边界点。其缺点是计算较繁琐,由于采用全局绝对密度参数,因此不能较好识别多密度数据集的聚类边界。BOURN 算法[16]是根据数据分布的统计

李光兴(1956—),男,硕士,副教授,主要研究方向为人工智能与数据挖掘,E-mail:lgx-22@163.com。

信息来识别边界模式的算法，由数据分布的均值和方差定义数据对象的边界度，按边界度的大小对数据集降序排列，选取前 t 个边界度最大的对象作为边界点输出。虽然该算法能有效地识别出边界点，但绝对参数选择较困难。文献[17-19]是基于网格的边界点识别算法，提出了利用相邻网格单元间的数据分布关系识别边界单元和边界点。算法的不足是运算较复杂。由于边界识别过程中以单元为单位，把边界网格中的数据点都作为边界点，识别出的边界不够精细，并且网格方法也不利于高维数据的处理。

另外，现有的文献很少讨论孤立点与边界点的关系，一般把孤立点与边界点的识别当成两个完全不同的过程。而实际上孤立点与边界点都具有其属性值明显不同于某些邻近对象的共性，又具有属性值偏离程度的差异，因而可以把孤立点与边界点的识别整合成一个过程。

针对现有的孤立点和边界点识别算法计算繁琐、难以对数据集的分布作预估、参数选取困难和处理多密度数据集功能不强，以及未对孤立点与边界点的识别整合处理等问题，提出了基于相对密度的孤立点和边界点的识别算法 (Recognition Algorithm of Outlier and Boundary Points Based on the Relative Density, OBRD)。

2 相关概念

$A_1 \times A_2 \times \cdots \times A_m$ 是一个 m 维数据空间，其中 $A_j (j=1,2,\cdots,m)$ 是第 j 个属性的有界定义域，$X=\{x_1, x_2, \cdots, x_n\}$ 是包含 n 个 m 维数据点的数据集，数据点 $x_i=(x_{i1}, x_{i2}, \cdots, x_{im})$ $(i=1,2,\cdots,n)$，其中 $x_{ij} \in A_j$。

定义 1　以数据点 $x_o=(x_{o1}, x_{o2}, \cdots, x_{om})$ 为圆心，以 r 为半径的区域称为 x_o 的 r 邻域，记为 $Nr(x_o)$。

定义 2　$Nr(x_o)$ 被超平面 $x=x_{oj}$ 分成关于 j 维的左邻域 $N_r^{-j}(x_o)$ 和右邻域 $N_r^{+j}(x_o)$，其中 $N_r^{-j}(x_o)$ 内的点 x_u 的第 j 维分量满足 $x_{uj} \leqslant x_{oj}$，$N_r^{+j}(x_o)$ 内的点 x_v 的第 j 维分量满足 $x_{vj} > x_{oj}$。

按定义，$N_r^{-j}(x_o)$ 包含了在 $Nr(x_o)$ 内且在超平面 $x=x_{oj}$ 上的点，$N_r^{+j}(x_o)$ 不包含超平面 $x=x_{oj}$ 上的点。

图 1 中两圆形区域分别是二维数据 x_p 和 x_q 的 r 邻域，x_p 的 r 邻域被垂直 I 维且经过 x_p 的铅直线分成了两部分，其中阴影部分为 $N_r^{-I}(x_p)$，另一部分为 $N_r^{+I}(x_p)$。x_q 的 r 邻域被垂直 II 维且经过 x_q 的铅直线分成了两部分，其中阴影部分为 $N_r^{-II}(x_q)$，另一部分为 $N_r^{+II}(x_q)$。

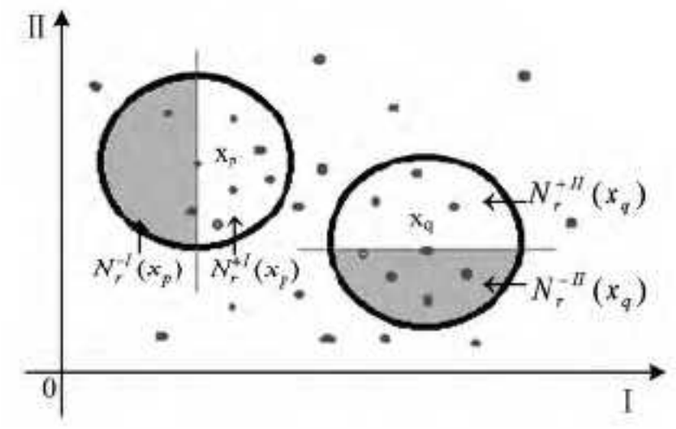


图 1　数据 x_p 和 x_q 的 r 邻域图

定义 3　$Nr(x_o)$ 中数据点的个数称为 $Nr(x_o)$ 的密度，记为 $Dr(x_o)$。$N_r^{-j}(x_o)$ 中数据点的个数称为 $N_r^{-j}(x_o)$ 的密度，记为 $D_r^{-j}(x_o)$。$N_r^{+j}(x_o)$ 中数据点的个数称为 $N_r^{+j}(x_o)$ 的密度，记为 $D_r^{+j}(x_o)$。

定义 4　m 维数据点 x_o 的孤立度 $IS(x_o)$ 定义为 x_o 的各维 r 左右邻域密度大者中的最小者与 $Dr(x_o)$ 的商。设 $a_j=\max(D_r^{-j}(x_o), D_r^{+j}(x_o))(j=1,2,\cdots,m)$，则 $IS(x_o)=\min(a_1, a_2, \cdots, a_m)/Dr(x_o)$。

$IS(x_o)$ 越大，表明 x_o 各维的 r 左右邻域密度相差越大，孤立度越大。孤立度的范围为 $[0,1]$。

定义 5　对给定的阈值 $\mu(0 \leqslant \mu \leqslant 1)$，若 $IS(x_o) \geqslant \mu$，则称 x_o 是孤立点。

显然，当 $Nr(x_o)$ 中除 x_o 外没有其它点时，有 $D_r^{-j}(x_o)=Dr(x_o)(j=1,2,\cdots,m)$，$I(x_o)=1$，大于或等于阈值的上限，$x_o$ 是孤立点。

定义 6　对 m 维数据点 x_o 的边界度 $B(x_o)$ 定义为 x_o 的各维 r 左右邻域密度中的最大者与 $Dr(x_o)$ 的商。设 $a_j=\max(D_r^{-j}(x_o), D_r^{+j}(x_o))(j=1,2,\cdots,m)$，则 $B(x_o)=\max(a_1, a_2, \cdots, a_m)/Dr(x_o)$。

$B(x_o)$ 越大，表明 x_o 某维的 r 左右邻域密度相差越大，因而 x_o 处于不同密度数据区域边缘的可能性越大。边界度的范围为 $[0,1]$。

定义 7　对给定的阈值 $\lambda(0 \leqslant \lambda \leqslant 1)$，若 x_o 不是孤立点，且有 $B(x_o) \geqslant \lambda$，则称 x_o 是边界点。

从几何上看，边界点为远离一些数据点，又与另一些数据点靠得较近，是位于不同密度数据区域边缘的数据对象。从属性上看，边界点就是在数据集合中与一些数据的某些属性差异较大，与另一些数据的某些属性差异较小的数据。

从几何上看，孤立点是远离数据集合中大多数数据的数据点。从属性来看，孤立点就是在数据集合中与大多数数据的某些属性相差较大的数据。

孤立点与边界点之间有一定联系，如果在定义 7 中取消 x_o 不是孤立点的限制，对给定的阈值 $\lambda(0 \leqslant \lambda \leqslant 1)$，若有 $IS(x_o) \geqslant \lambda$，则一定有 $B(x_o) \geqslant \lambda$；若有 $B(x_o) \geqslant \lambda$，则不一定有 $IS(x_o) \geqslant \lambda$。由此可见，孤立点也可看成特殊的边界点，反之则不一定成立。

3 OBRD 孤立点和边界点识别算法

3.1 OBRD 算法步骤

OBRD 算法包括查找近邻居、统计 x_o 的 r 邻域及其各维左右邻域的密度、确定边界度和孤立度，判断边界点和孤立点等过程。具体步骤如下。

输入：数据集 X，近邻半径 r，孤立点阈值 μ，边界点阈值 λ

输出：孤立点和边界点

步骤 1　从任意一个未判断的数据点 x_o 出发，查找 x_o 的所有距离小于 r 的邻居。若所有的数据点都作出判断，则执行步骤 5，否则执行下一步。

步骤 2　统计 $Nr(x_o)$ 的密度 $Dr(x_o)$，以及 x_o 每一维的左右邻域密度 $D_r^{-j}(x_o)$ 和 $D_r^{+j}(x_o)(j=1,2,\cdots,m)$。

步骤 3　计算孤立度 $IS(x_o)$，结合给定的阈值判断数据 x_o 是否是孤立点。若是执行步骤 1，否则执行下一步。

步骤 4　计算边界度 $B(x_o)$，结合给定的阈值判断数据 x_o 是否是边界点，然后执行步骤 1。

步骤 5　输出孤立点和边界点。算法结束。

经实验，近邻半径 r 一般不大于数据集中数据的各维分量全距中的最小者的 1/10，且使任意一个数据点的 r 邻域密度不超过数据集中数据个数的 1/4。孤立点阈值和边界阈值一般在 $[0.6, 0.9]$ 范围内选取，并且孤立点阈值不小于边界阈值。

3.2 算法复杂度

m 维数据空间中有 n 个数据，按近邻半径 r 取不大于数据集中数据的各维分量全距中最小者的 $1/10$，一个数据的所有距离小于 r 的邻居数一般最多有 $0.5n$ 个，算法的时间复杂度为 $O(0.5n^2)$。

4 实验结果及分析

实验环境是 CPU 为 2×1.67GHz，操作系统为 Win7，数据集由 MATLAB 随机函数 rand 产生，程序用 MATLAB 实现。

4.1 算法性能实验

OBRD 实验参数为近邻距离 $r=10$，孤立点阈值和边界阈值都为 0.7。BORDER 实验参数为近邻居数 10，在相同数据规模情况下，取 BORDER 边界点数量与 OBRD 算法识别的边界点数相同。

数据规模实验采用 9 个二维数据集，分布呈正六形，数据个数分别为：11174，20541，30006，38213，50262，67256，83662，104922，128078。随着数据量递增，OBRD 与 BORDER 算法的执行时间差距加大（见图 2），OBRD 比 BORDER 算法的执行效率更高。原因是 OBRD 算法判别某个数据点是否是边界点，只与该数据点的 r 邻域有关，与其它点的 r 邻域无关，即查找几何范围是局部。而 BORDER 算法确定一个对象是多少个其它对象的反向 k 近邻，查找几何范围不确定，通常需要遍历整个数据集，大大增加算法的时间花销。

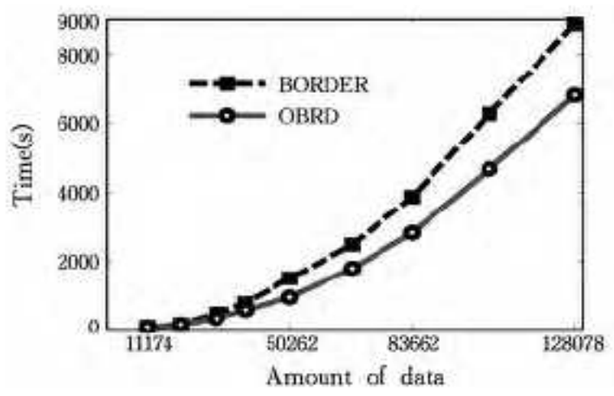


图 2 算法数据规模实验的执行时间比较

维数扩展实验采用数据集包含 101479 个数据，维数从 3 维扩展到 12 维。从图 3 可见随着维数的增加，OBRD 与 BORDER 算法的执行时间都成线性增加，表明 OBRD 算法具有良好的维数可扩展性。

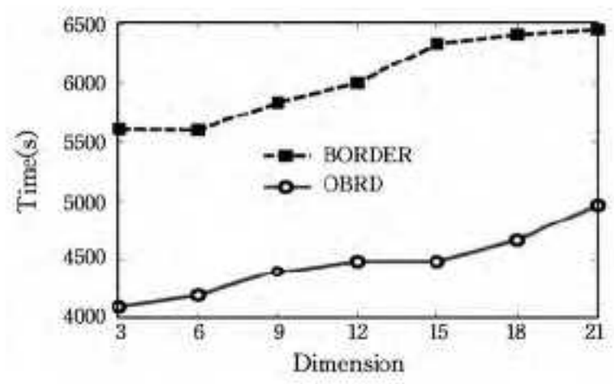


图 3 算法维数扩展实验的执行时间比较

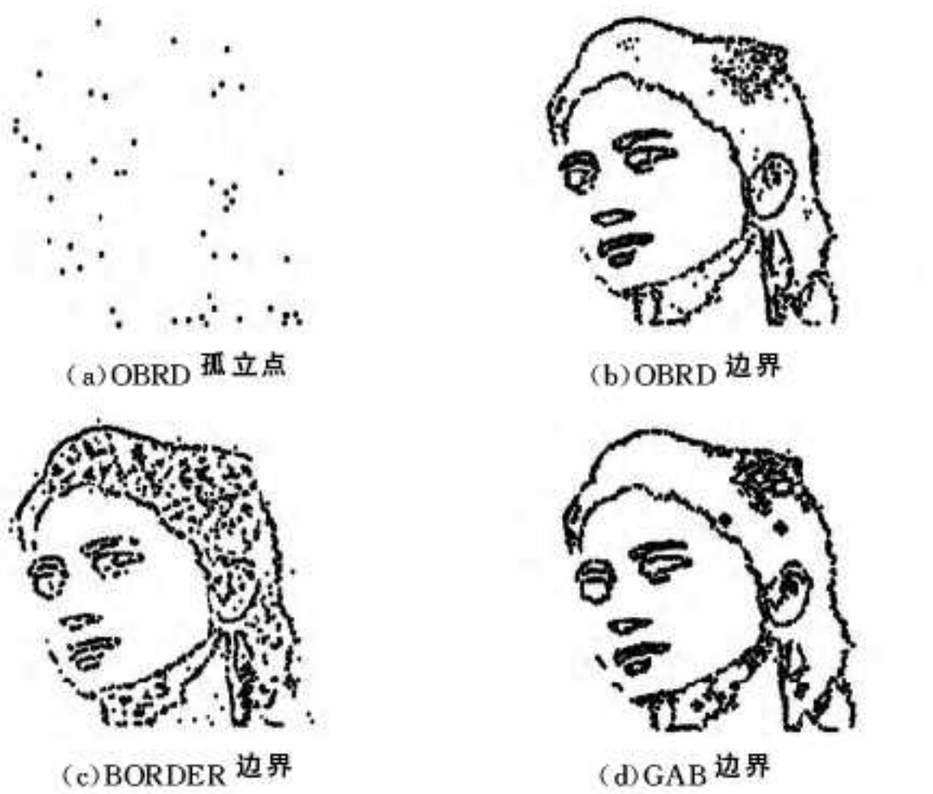
4.2 算法有效性对比实验

实验 1 采用包含 13908 个数据的二维多密度头像数据集（见图 4(a)）。头像数据集的数据的第 I 维和第 II 维分量全距分别是 476 和 561。构成头像的头发、眼、鼻、嘴、颈部及附近区域的密度不同，形状各异。实验 2 采用包含 12348 个数据的二维多密度盆景数据集（见图 4(b)）。数据集中的数据的第 I 维和第 II 维分量全距分别是 327.22 和 258.31。构成盆景的叶、枝、盆及地面上的散点区域等部分密度不同，形状零乱。

(a) 头像数据集　(b) 盆景数据集

图 4 原始数据分布图

实验 1 取 OBRD 近邻半径 $r=10$，孤立点阈值为 0.82，边界阈值为 0.71。识别出 54 个孤立点（见图 5(a)），2540 个边界点（见图 5(b)）。BORDER 参数为近邻个数 30，边界点 2540 个（见图 5(c)）。GAB 算法[17]是基于网格的边界点识别算法，具有去掉噪声功能。取网格划分数为 83，边界阈值为 0.08，识别边界点 3082 个（见图 5(d)）。



(a) OBRD 孤立点　(b) OBRD 边界

(c) BORDER 边界　(d) GAB 边界

图 5 头像数据集边界识别效果比较

实验 2 取 OBRD 近邻半径 $r=4$，孤立点阈值为 0.82，边界阈值 0.7。识别出 96 个孤立点（见图 6(a)），2761 个边界点（见图 6(b)）。BORDER 参数为近邻个数 30，识别边界点 2761 个（见图 6(c)）。GAB 算法[19]取网格划分数为 78，边界阈值 0.11，识别边界点 2949（见图 6(d)）。

(a) OBRD 孤立点　(b) OBRD 边界

(c) BORDER 边界　(d) GAB 边界

图 6 盆景数据集边界识别效果比较

从这两个实验的效果来看，OBRD 识别的外围边界在完整性方面优于 BORDER 和 GAB，如 BORDER 识别的头像集

（下转第 280 页）

tual time-based network emulation testbed[J]. Journal of Simulation,2014,8(3):206-214

[6] Mustafee, Navonil, Taylor, et al. High-performance simulation and simulation methodologies[J]. Journal of Simulation, 2013, 89(11):1291-1292

[7] 刘畅,刘西洋,陈平. 大规模网络仿真方法研究[J]. 计算机仿真, 2005,22(5):8-11,15

[8] 李广俊. 一种分布式网络仿真系统的设计与实现[D]. 成都:电子科技大学,2007

[9] 韦涛,田永春,姜永广. 基于协同的半实物网络仿真系统设计[J]. 通信技术,2011,12(44):64-66

[10] 王国玉,冯润明,陈永光. 无边界靶场——电子信息系统一体化联合试验评估体系与集成方法[M]. 北京:国防工业出版社, 2007

[11] 李洋. 网络协议本质论[M]. 北京:电子工业出版社,2011

[12] 毕亿默,卢超,王华. 一种数据交换整合平台的设计与实现[J]. 计算机应用与软件,2013,30(12):127-129,136

[13] 维克托. 迈尔—舍恩伯格. 删除:大数据取舍之道[M]. 杭州:浙江人民出版社,2013

[14] 维克托. 迈尔—舍恩伯格. 大数据时代[M]. 杭州:浙江人民出版社,2013

[15] 纳特. 西尔弗. 信号与噪声[M]. 北京:中信出版社,2013

[16] 黎富贵,吴宙华. 基于 Web Services 动态资源服务化的研究[J]. 计算机应用与软件,2014,31(9):83-85,165

(上接第 238 页)

右前额上的头发部分边界欠缺,GAB 识别的盆景植物主干部分边界欠缺等。OBRD 是逐点识别边界点的算法,因而比 GAB 以单元为单位识别边界点更精细,如头像集中头发上的花夹和耳坠边界、盆景数据集中枝叶边界等。由于 BORDER 把处于聚类内部的很多点也当作边界点,如形成头发内部的一些点,因此边界识别效果较差。

4.3 参数敏感性实验

对 OBRD 算法中近邻半径 r,孤立点阈值 μ 和边界点阈值 λ 3 个参数进行敏感性实验。数据集采用 4.2 节实验 1 中的头像数据集。设置 $\mu=0.82,\lambda=0.71$ 时,r 从 16 逐次减少 1 直到 7,当 r 每减少 1 时,孤立点数平均增加 4.64%,边界点数平均增加 2%,可见 OBRD 算法对近邻半径 r 不是很敏感。$r=10,\lambda=0.71$,μ 从 0.87 逐次减少 0.01 直到 0.78,识别的孤立点数增加 9.64%;$r=10,\mu=0.82$,λ 从 0.76 逐次减少 0.01 直到 0.67,当 λ 每减少 0.01 时,识别的边界点数平均增加 7.55%,可见 OBRD 算法对孤立点阈值和边界阈值较敏感。

结束语 利用数据点的 r 领域和半领域的密度相对数能反映数据点所在局部区域数据分布差异的特征,给出了相应孤立点和聚类边界点识别算法 OBRD。由于一个数据点的 OBRD 识别只与它的 r 邻域的密度有关,因此算法具有简洁、执行效率较高、与数据输入顺序无关等特点,孤立点和边界点可以在同一个过程中识别出来。理论分析和实验显示,OBRD 算法能识别出呈任意形状分布的多密度数据集的孤立点和聚类边界点,能够识别出紧挨边界的孤立点,且输入参数少,具有较高的识别精度,但参数的敏感性还需进一步改进。

参 考 文 献

[1] Branch J W, Giannella C, Szymanski B, et al. In-network outlier detection in wireless sensor networks[J]. Knowledge and Information Systems, 2013, 34(1):23-54

[2] 贾润达,刘俊豪,毛志忠,等. 基于鲁棒 M 估计的间歇过程离群点检测[J]. 仪器仪表学报,2013,34(8):1726-1729

[3] 黄毅群,卢正鼎,胡和平,等. 分布式异常识别中隐私保持问题研究[J]. 电子学报,2006,34(5):796-799

[4] Niu Z, Shi S, Sun J, et al. A survey of outlier detection methodologies and their applications[M]// Artificial Intelligence and Computational Intelligence. Springer Berlin Heidelberg, 2011: 380-387

[5] Zoubi M B A, Obeid N. A Fast Distance Algorithm to Detect Outliers[J]. Journal of Computer Science, 2007, 3(12):944-947

[6] Zhang Yue, Yang Xue-hua, Li Huang. An Outlie Mining Algorithm Based on Confidence Interval[C]// Proc. of the 2nd IEEE International Conference on Information Management and Engineering. IEEE Press, 2010

[7] Bhaduri K, Matthews B L, Giannella C R. Algorithms for speeding up distance-based outlier detection[C]// Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2011:859-867

[8] Keller F, Müller E, Böhm K. HiCS: high contrast subspaces for density-based outlier ranking[C]// 2012 IEEE 28th International Conference on Data Engineering (ICDE). IEEE, 2012:1037-1048

[9] Aggarwal C C, Philip S Y. Outlier Detection with Uncertain Data[C]// SDM. 2008:483-493

[10] 李光兴,杨燕. 基于网格相邻关系的离异点识别算法[J]. 计算机工程与科学,2010,32(9):130-133

[11] 赵峰,秦锋. 基于单元的孤立点识别算法改进及应用[J]. 计算机工程,2009,35(19):78-80

[12] 张选平,祝兴昌,马琮. 一种基于边界识别的聚类算法[J]. 西安交通大学学报,2007,41(12):1387-1390

[13] 楼晓俊,孙雨轩,刘海涛. 聚类边界过采样不平衡数据分类方法[J]. 浙江大学学报(工学版),2013,47(6):944-949

[14] Xia C, Hsu W, Lee M L, et al. BORDER: efficient computation of boundary points[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(3):289-303

[15] 吾守尔·斯拉木,李丰军,陶梅. IBORA:一种改进的有效的边界点检测[J]. 小型微型计算机系统,2008,29(10):1845-1848

[16] 邱保志,张枫,岳峰. 基于统计信息的聚类边界模式识别算法[J]. 计算机工程,2008,34(3):91-93

[17] Li G, Li B. Boundary Point Recognition Algorithm Based on Grid Adjacency Relation[M]// Recent Advances in Computer Science and Information Engineering. Springer Berlin Heidelberg, 2012:211-218

[18] 张鸿雁,刘希玉,付萍. 一种网格聚类的边缘识别算法[J]. 控制与决策,2011,26(12):1846-1850

[19] 邱保志,余田. 基于网格梯度的边界点识别算法的研究[J]. 微电子学与计算机,2008,25(3):77-80