# 基于平均密度的孤立点检测研究

施化吉[1,2]，周书勇[1]，李星毅[1,3]，唐 慧[1]，丁秋林[2]

(1. 江苏大学计算机科学与通信工程学院 江苏 镇江 212013; 2. 南京航空航天大学计算机应用研究所 南京 210016;
3. 北京交通大学电子信息工程学院 北京 海淀区 100044)

【摘要】为了使孤立点检测更为自动化，减少用户对参数选择的困难，提出了平均密度的定义，并给出基于平均密度的孤立点检测方法。该方法提出了孤立点对象的密度要小于数据集的平均密度；非孤立点对象的密度不应因为封闭区间的收缩而减少。采用企鹅图像边缘检测对该方法进行验证，实验结果表明，该方法能够有效地检测出图像边缘孤立点，同时简化了孤立点检测时对用户输入参数的要求。

关 键 词 平均密度; 平均距离; 对象密度; 孤立点检测
中图分类号 TP311 文献标识码 A

# Average Density-Based Outliers Detection

SHI Hua-ji[1,2]，ZHOU Shu-yong[1]，LI Xing-yi[1,3]，TANG Hui[1]，DING Qiu-lin[2]

(1. School of Computer Science and Telecommunication Engineering, Jiangsu University Zhenjiang Jiangsu 212013;
2. Computer Application Institute, Nanjing University of Aeronautics and Astronautics Nanjing 210016;
3. School of Electronics and Information Engineering, Beijing Jiaotong University Haidian Beijing 100044)

**Abstract** In order to make the outlier detection more automatic and decrease the users' difficulty for the selection of parameters, an outlier detection method with a new definition of average density is proposed. In this method, the outlier's density is considered smaller than the average density of data set and the none-outlier's density shouldn't decrease with its closed interval compression. An experiment is used to identify the outline of the animal's body. The experimental results show that the method identifies the face's outline effectively.

**Key words** average density; average distance; object density; outlier detection

孤立点检测是数据挖掘的一个重要方面[1]，近年来受到越来越多的重视。其任务是用来发现数据集中小的模式，即数据集中明显不同于其他数据的对象。当前研究的热点主要是关注孤立点的应用驱动，如信用卡欺骗、入侵检测、气象预报、公共卫生、医疗等。

文献[2]给出了孤立点的本质性定义：孤立点是一个观测值，它与其他的点是如此的不同，以至于怀疑它是产生于完全不同的机制。后来研究者们根据对异常存在的不同假设，发展了很多孤立点检测算法，大体上可分为基于统计的、基于距离的、基于密度的、基于聚类的等。但这些算法在自动化上都存在不足，即都要求用户输入必要的参数，算法对参数的依赖性较强。而参数的选择通常比较困难，要求用户具备丰富的经验，并且需要多次反复才能达到效果，对用户的使用提出了较高的要求。在分析传统算法的基础上，本文提出了平均密度的定义，在平均密度概念下的孤立点，其含义更加接近文献[2]的孤立点本质性定义，更符合人们对孤立点的认识；在通常的孤立点检测时，不依赖于用户对阈值设置的要求，使算法更加自动化。

## 1 相关研究

传统的孤立点检测方法大致分为四类：基于统计的方法[3]、基于距离的方法、基于密度的方法和基于聚类的方法。

(1) 基于统计的方法：主要思想是假定数据集服从某种分布或概率模型，通过不一致检验把那些严重偏离分布曲线的记录视为孤立点。对于单个属性，存在各种统计孤立点检测，检测效果较好，然而到了二维以上，检测效果会变差。

首先，此法检测出来的孤立点很可能被不同的

收稿时间：2007 – 09 – 07
基金项目：国家火炬计划项目(2004EB33006)
作者简介：施化吉(1964 – )，男，博士生，教授，主要从事数据库与数据挖掘、信息安全等方面的研究.

分布模型检测出来，因此产生这些孤立点的机制可能不唯一，对孤立点的解释性不足；其次，基于统计的方法在很大程度上依赖于待挖掘的数据集是否满足某种概率分布模型，模型的参数、离群点的数目等对基于统计的方法都有非常重要的意义，而确定这些参数通常都比较困难。

(2) 基于距离的方法：该方法最早是由文献[4-6]提出的，即把记录看作高维空间中的点，孤立点被定义为数据集中与大多数点之间的距离都大于某个阈值的点，通常被描述为DB(pct,$d_{\min}$)，数据集$T$中一个记录$O$称为孤立点，当且仅当数据集$T$中至少有pct部分的数据与$O$的距离大于$d_{\min}$。该方法认为一个对象是孤立点，它必远离大部分对象。

这种方法使用的是全局阈值，它不能处理具有不同密度的数据集。此外，算法需要事先确定参数pct和$d_{\min}$，这是比较困难的，特别是对不同聚类密度数据集而言，其参数$d_{\min}$会有很大差异，并且一般无规律可循。因此，对于给定的不同参数$d_{\min}$，孤立点检测结果通常具有很大的不稳定性。

(3) 基于密度的方法[6-7]：方法中的密度通常用邻近度定义，如定义密度为到$k$个最近邻的平均距离的倒数。如果该距离小，则得分高。该方法认为：孤立点是在低密度区域中的对象。基于密度的方法，给出了对象是孤立点程度的定量度量，并且即使数据具有不同密度的区域也能很好地处理。但这些方法必然具有$O(n^2)$的时间复杂度，其参数选择也是困难的。

(4) 基于聚类的方法[10]：首先聚类所有对象，然后评估各对象属于簇的程度，或把远离其他族的小簇视为孤立点。此方法中，有些聚类的时空复杂度是线性或近于线性的，因而它们可能是高效的。但所产生的孤立点集也会非常依赖于所用的簇的个数$k$和数据中孤立点的存在性。聚类算法产生的簇的质量对该算法产生孤立点的质量影响很大。

## 2　基于平均密度的孤立点检测

### 2.1　平均密度概念

本文定义的“平均密度”类似于物理学上的平均密度，其定义如下：设数据集$T$的维度为$m$，对象的个数为$c$，数据集中所有对象间最大距离为$d$，则平均密度为：

$$S=\frac{c}{t\left(\frac{\sqrt{3}}{2}\times d\right)^m} \tag{1}$$

式中　$t$为常系数，若设为1，则公式右边的分母表示对象所属多维空间的超体积。

为了便于说明，本文用二维数据对象来阐述，如图1所示，设在平面上的$n$个点，找出其中距离最大的两点$d_1$和$d_2$，以$d_1d_2$为直径分别向两端延长至原直径的$\sqrt{3}$倍，这样可保证所表示的圆封闭区间能包括所有的点。

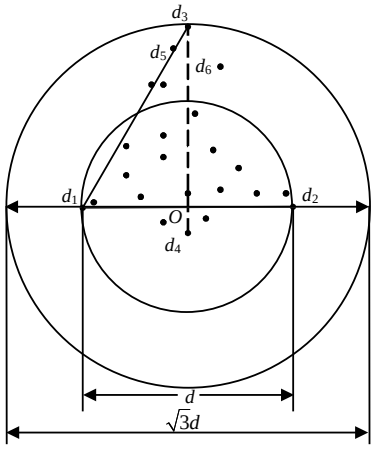



图1　二维空间平均密度示意图

假设只以图中距离最远的两点$d_1$和$d_2$之间的长度为直径画圆，所画出的封闭区间不能包含所有的点，像$d_3$、$d_5$、$d_6$点就落到区间之外，这样计算的平均密度可能会因遗漏一些点而不能反映客观情况。而适当扩大范围之后的封闭区间则能包含所有的对象点，且扩大后的直径是原来的$\sqrt{3}$倍，从几何学上恰好能够保证所绘制封闭区间包含所有数据集中的对象点。证明如下：

如图1所示，设内圆的半径为1，则图中两点间最长距离$\overline{d_1d_2}$=2，如果$\overline{d_1d_3}$=2，则由直角三角形$\Delta d_1Od_3$知$\overline{Od_3}=\sqrt{3}$，即外圆的半径$Od_3$是内圆半径$Od_1$的$\sqrt{3}$倍。

以上二维空间的平均密度的概念同样适用于三维或多维空间。

### 2.2　基于平均密度的方法设计

如上所述，本文对传统的孤立点算法的密度定义作了改进，主要引入对象的维度$m$作为对象间距离$r$ ($r$=$D$/2)的幂。相应的孤立点检测方法如下：

本文仍沿用基于密度的孤立点检测算法思想：孤立点是在低密度区域中的对象，且认为一个数据集中若存在孤立点，那么孤立点的密度会小于平均密度。接下来定义每个对象密度的计算方法，对原始数据集进行标准化后，计算$n$个对象两两之间的距离$d_{i,j}$，形成距离矩阵$\boldsymbol{R}$：

$$\boldsymbol{R}=\begin{bmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,n} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ d_{n,1} & d_{n,2} & \cdots & d_{n,n} \end{bmatrix} \tag{2}$$

设所有对象的平均距离为：

$$D=\frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d_{i,j}}{C_n^2} \tag{3}$$

为了与平均密度的定义一致，本文在计算每个对象$i$的对象密度时，仍将有效直径延长至原来的$\sqrt{3}$倍，即以$\sqrt{3}D$为单位封闭区间的直径，扫描数据矩阵$\boldsymbol{R}$，计算每个对象单位封闭区间的对象数$C_i$，即与对象$i$间距离小于$\sqrt{3}D/2$的对象点计数，因此每个对象$i$的对象密度为：

$$S_i=\frac{c_i}{\left(\frac{\sqrt{3}}{2}D\right)^m} \tag{4}$$

### 2.3 基于平均密度的孤立点检测

有了平均密度$S$和对象密度$S_i$后，只需比较这两个密度就能检测出孤立点：

$$\begin{cases} S_i \geq S & i\text{对象非孤立点} \\ S_i < S & i\text{对象是孤立点} \end{cases} \tag{5}$$

但是，这样检测孤立点可能会存在问题，因为本文在假设封闭区间时，把直径延长了，可能会淹没掉许多内部的孤立点。即许多对象尽管其本身密度并不高，但可能被许多高密度的对象远远的包围着，而不能被检测出来，如图2所示，

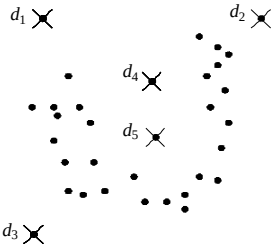



图2 数据集内部孤立点的检测

单纯的密度比较如式(5)，可能只能检测出数据集中外围的孤立点如$d_1$、$d_2$、$d_3$，而同样稀疏的$d_4$、$d_5$可能就被淹没了。

为了解决这类问题，本文提出了基于平均密度方法思想的另一个方面：非孤立点的密度不因半径的收缩而减少太多，因此可以进行“收缩计算”。所谓收缩计算是指在扫描计算$C_i$的同时。计算该对象$i$的另一个收缩了的封闭区间，不妨设该封闭区间的直径为$aD$($a$为系数，通常小于1，随数据对象的维度及处理的数据量而定)，在这样的封闭区间下重新计算对象$i$ 的对象数$C_i'$，再重新计算对象$i$的对象密度为：

$$S_i'=\frac{c_i'}{\left(\frac{\sqrt{3}}{2}aD\right)^m} \tag{6}$$

所以对于对象$i$，如果只通过式(5)未能判定其是孤立点，则需通过式(7)来继续判断其是否为孤立点：

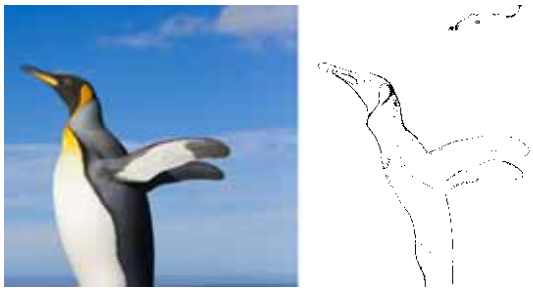
$$\begin{cases} S_i' \geq rS_i & i\text{对象非孤立点} \\ S_i' < rS_i & i\text{对象是孤立点} \end{cases} \tag{7}$$

式中 $r$为系数，通常小于1。

通过以上两次判断不仅能检测出数据集中外部的孤立点，也能检测出内部的孤立点，对于密度不均匀的数据也有较好的检测效果。

## 3 实验分析

本文选择了一幅动物(企鹅)彩色图像进行实验，图像为BMP格式文件，图像的色彩是24位，大小300×300，总共包含90 000个像素点。实验目的是对企鹅的轮廓进行提取[11]。实验原理：孤立点检测是发现数据集中小的模式，也就是数据集中被认为与其他数据不相似或不一致的数据对象，而企鹅图像边缘的像素值与其他像素值有明显的差别。由此，可以考虑将边缘的像素作为孤立点，通过本文所提的算法实现企鹅的边缘检测。在具体的实现过程中，取位图的RGB值作为数据集，然后利用上述算法，对这些色彩值进行孤立点发现，最终将得到的孤立点重新绘出。孤立点检测及原始图如图3所示。



a. 原始图 b. 孤立点检测图

图3 24位真彩色图的立点检测

考虑到数据集较大，本实验在计算平均密度$S_i$和平均距离$D$时，采用了隔行抽样，再逐行扫描位图的像素值进行孤立点的检测。这样可能损失了精度，但极大地降低了数据处理的规模。且实验结果表明该处理方法可行、有效。

(下转第1295页)

[2] 赵永恒. fits文件解析[EB/OL]. www.lamost.org, 2007-02-19.
[3] 覃冬梅. 一种基于主分量分析的恒星光谱快速分类法[J]. 光谱学与光谱分析, 2003, 23(1): 182-186.
[4] 李乡儒. 几个学习算法及其在星系光谱分类中的应用[D]. 北京: 中国科学院北京天文台，2007.
[5] 边肇祺, 张学工. 模式识别[M]. 北京: 清华大学出版社, 2000.
[6] 苏金明. Matlab工具箱应用[M]. 北京: 电子工业出版社, 2004.
[7] 薛建桥. 神经网络技术与光谱自动分类[D]. 北京: 中国科学院北京天文台, 1999.
[8] KURTZ M J. Progress in automation techniques for mk classfication[J]. Astrophys, 2004, (4): 111-117.
[9] CHEN OT-C. Motion estimation using a one-dimentional gradient descent search[J]. IEEE Transactions Circuits and System for VideoTechnology, 2000, 10(4): 608-616.
[10] BAILER-JONEA CAL. Techniques for mk classfication[J]. Astrophysics and Space Science, 2002, (24): 21-30.
[11] 王文胜. 图像特征抽取的奇异值分解方法[J]. 计算机工程, 2006, 32(8): 32-36.

编　辑　漆　蓉

---

(上接第1288页)

## 4　结　论

本文介绍了孤立点检测的传统算法，并在此基础上，提出了平均密度的概念，平均密度接近物理学上的关于密度的定义，使人们对孤立点的认识更自然；在平均密度概念的基础上，给出了基于平均密度的孤立点检测方法，该方法对孤立点的检测更加自动化，通常情况下，它不依赖于用户输入参数。

和基于密度的或基于距离的大多数孤立点检测算法一样，该方法的时间复杂度是$O(n^2)$，在数据规模较大时，需考虑抽样来确定平均密度$S_i$和平均距离$D$，再对各数据对象进行孤立点检测。本文在传统孤立点定义的基础上，拓展了新的视点，在算法自动化上作了一定的探索。

### 参　考　文　献

[1] HAN J, KAMBER M. Data mining: concepts and techniques[M]. [S.l.]: Morgan Kaufmann Publishers, Inc. 2001.
[2] HAWKINS D M. Identification of outliers[M]. London: Chapman and Hall, 1980.
[3] BARNETT V, LEWIS T. Outliers in statistical data[M]. New York: John Wiley & Sons, 1994.
[4] KNORR E M, NG RT. Algorithms for mining distance-based outliers in large datasets[C]//In: Proceedings of the 24th VLDB Conference. New York: [s.n.], 1998: 392-403.
[5] RAMASWAMY S, RASTOGI R, SHIM K. Efficient algorithms for mining outliers from large data sets[C]//In: Proceedings of the ACM SIGMOD Conference. [S.l.]: [s.n.], 2000: 473-438.
[6] 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): 1009-1015.
[7] BREUNIG M M, KRIEGEL H, NG R T, et al. LOF: Identifying density-based local outliers[C]//In: Proc of the 2000 ACM SIGMOD Int'1 Conf on Management of Data. Dallas: ACM Press, 2000: 93-104.
[8] PAPADIMITIROU S, KITAGAWA H, GIBBONS P B, et al. LOCI: Fast outlier detection using the local correlation integral[C]//In: Proc of the 19th Int'1 Conf on Data Engineering. [S. l.]: IEEE Computer Society Press, 2003.
[9] 蒋盛益, 李庆华, 王　卉, 等. 一种增强的局部异常挖掘方法[J]. 计算机研究与发展, 2005, 42(2): 210-216.
[10] HARDIN J, ROCKE D M. Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator[J]. Computational Statistics and Data Analysis, 2004, 44: 625-638.
[11] 邵峰晶, 孙仁诚, 郭振波. 基于孤立点发现的彩色图像人脸边缘提取算法[J]. 计算机科学, 2006, 33(9): 201-203.

编　辑　张　俊