DOI: 10.11992/tis.201809032
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20181229.1113.002.html

基于自然邻居邻域图的无参数离群检测算法

冯骥, 冉瑞生, 魏延
(重庆师范大学 计算机与信息科学学院, 重庆 401331)

摘 要: 数据挖掘领域, 基于最近邻居思想的离群检测算法在面对复杂数据时, 很难在没有足够先验知识条件下进行适当的参数选择。为了解决这个问题, 本文在自然邻居方法的基础上, 提出一种利用加权自然邻居邻域图进行离群检测的算法。该算法在整个过程不需要人为设置参数, 并且能在不同分布特征的数据中准确找到数据集中的全局离群点和局部离群点。人工数据集和真实数据的离群检测结果均证明, 本算法能够取得和有参数的算法中最优参数相近的效果, 算法检测结果远好于对参数敏感算法的大部分情况, 且更优于对参数不敏感的算法, 具有更强的普适性和实用性。

关键词: 无参数; 自适应; 最近邻居; 加权图; 离群检测; 离群因子; 全局离群点; 局部离群点
中图分类号: TP311 **文献标志码**: A **文章编号**: 1673−4785(2019)05−0998−09

中文引用格式: 冯骥, 冉瑞生, 魏延. 基于自然邻居邻域图的无参数离群检测算法 [J]. 智能系统学报, 2019, 14(5): 998–1006.
英文引用格式: FENG Ji, RAN Ruisheng, WEI Yan. A parameter-free outlier detection algorithm based on natural neighborhood graph[J]. CAAI transactions on intelligent systems, 2019, 14(5): 998–1006.

A parameter-free outlier detection algorithm based on natural neighborhood graph

FENG Ji, RAN Ruisheng, WEI Yan
(College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China)

Abstract: This study aims to deal with the practical shortages of nearest-neighbor-based data mining techniques, particularly outlier detection. In particular, when data sets have arbitrarily shaped clusters and varying density, determining the appropriate parameters without a priori knowledge becomes difficult. To address this issue, on the basis of the natural neighbor method, which can better reflect the relationship between elements in a data set than the *k*-nearest neighbor method, we present a graph called the weighted natural neighborhood graph for outlier detection. The weighted natural neighborhood graph does not need to set parameters artificially in the entire process and can identify global and local outliers in the data set with different distribution characteristics. The outlier detection results of artificial dataset and real data prove that the algorithm can obtain an effect similar to that of the optimal parameter in the algorithm with parameters. The algorithm detection result is far better than that of most parameter-sensitive algorithms and is much better than that of the parameter-insensitive algorithm, which has stronger universality and more practicality.

Keywords: parameter-free; adaptive neighbor; nearest neighbor; weighted graph; outlier detection; outlier factor; global outlier; local outlier

随着大数据技术和数据密集型科学的发展, 数据已经渗透到各个行业和业务功能中, 成为了生产的一个重要因素。越来越多的国家、政府、行业、企业等机构已经意识到大数据正在成为组织最重要的资产, 数据分析能力也已经成为组织的核心竞争力。目前, 国家、政府已经把大数据

收稿日期: 2018−09−16. **网络出版日期**: 2019−01−04.
基金项目: 教育部人文社会科学研究项目 (18XJC880002); 重庆市教委科技项目 (KJQN201800539); 重庆市自然科学基金项目 (cstc2013jcyjA40049); 重庆师范大学基金项目 (17XLB003).
通信作者: 冯骥. E-mail: jifeng@cqnu.edu.cn.

应用推进了人们的生活中, 大数据研究也成为了“十三五”期间的重点发展项目。

离群检测也是大数据战略中举足轻重的核心技术, 在大数据技术发展日新月异之际, 包括离群检测在内, 数据挖掘中聚类、分类等技术也随之不断地进步与发展。离群检测的目的在于检测出数据集中那些被怀疑由异常机制产生的奇异数据, 广泛应用于欺诈检测[1]、异常检测[2]、图像检测[3]、医学分析[4]、信号异常检测[5] 等, 并取得了众多令人满意的结果。

然而迄今为止, 对离群的定义没有一个统一的认识, 因离群定义的不同, 离群检测算法的检测结果也会有所不同。本文将同时考虑全局离群点和局部离群点, 提出了基于加权自然邻居邻域图的离群检测算法 (weighted natural neighborhood graph outlier detection algorithm, WNaNG)。该算法利用加权自然邻居邻域图计算局部自然邻居离群度, 找出离群度最高的 n 个离群点, 而无需人为地预设邻域参数 k。在此基础上, 本算法可以利用离群度的离散图辅助挖掘出合理的离群点, 或利用长尾理论直接根据离群度的分布找到合理的离群点区间, 进而去除参数 n, 在无需邻域参数 k 的基础上将算法完善为完全无参数的离群检测算法。

1　传统离群检测算法

为了解决局部离群点的问题, 基于密度的离群检测算法被陆续提出 (例如 LOF 算法)。与距离度量不同的是, 基于密度的离群检测算法通过定义各种不同的局部离群度来检测局部离群点。局部离群度往往能够准确地反映出数据点与其周围点密度特性上的差异, 进而可以通过局部离群度的大小直观地找到离群点。这种方法在面对局部离群点时, 往往能够取得更好的检测效果。

近年来, 随着数据挖掘研究领域的不断深入, 针对不同的应用领域, 研究人员提出了大量的改进算法。Kim 等[6] 利用 k-d 树和近似 k-最近邻居方法提出了一种高效的离群检测算法; Campello 等[7] 则提出了一种基于层次密度的数据挖掘算法, 该算法提高了传统的基于密度的聚类效果, 并通过计算得到层次化的聚类结果进行聚类、离群检测和数据可视化等多项任务。苟和平等[8] 利用 DBSCAN(density-based spatial clustering of applications with noise) 算法对样本进行去噪和裁剪提出了 KNN(k-nearest neighbor) 算法的改进算法; 周芳芳等[9] 以体数据的标量值与梯度模直方图的密度分布为基础对传统算法进行了改进; 周国兵等[10] 基于算法空间复杂度的考虑提出了 Bit k-means 算法; 王习特等[11]提出了 BOD(BDSP-based outlier detection) 分布式离群点检测算法解决了传统的集中式算法处理效率受限的问题; 陆海青等[12] 提出的图像分割算法需根据图像噪声的强度适当地选取邻域窗口大小, 并根据邻域窗口中各像素的灰度差异, 利用指数函数进一步控制邻域像素的影响权重, 实现像素灰度的自适应加权, 从而提高像素灰度计算的准确性; 赵冠哲等[13] 提出的异常检测方法针对移动数据中历史位置和好友圈信息进行高效的检测, 并在检测方法中探讨了针对不同情况时邻域参数的选择策略; 张美琴等[14] 提出了一种基于加权聚类集成的标签传播算法, 该算法利用逆邻居的思想完成了计算基聚类集, 进而利用基聚类集的加权相似性矩阵得到社区划分结果。

上述算法在各自的应用领域中均取得了令人满意的效果, 进一步推动了相关技术的发展, 却也同时凸显出了另一个亟待解决的问题——邻域参数对算法效率的影响。大多数基于距离和基于密度的离群检测算法的核心都架构于 k-最近邻居思想, 因此邻域参数的选择也就是 k 值的设置。k 值较大会导致短路, 使得算法在离群点检测中错误地将部分局部离群点归到正常点的范围中, 而 k 值较小则会导致数据簇的不完整, 将边缘点与稀疏点错误地归为局部离群点。更为困难的是, 选择一个恰当具有非普适性的 k 值, 在一个数据集中合适的 k 值通常在其他数据集中都会成为一个不恰当的选择。解决邻域参数的选取问题出现了两种思考的方向: 探寻具有普适性的邻域参数选择方法, 或降低邻域参数的敏感性。

Ha 等[15] 利用不稳定因子提出了一种新的对参数不敏感的离群检测算法 INS(instability factor)。INS 改善了 KNN 算法中离群检测算法对参数敏感的问题, 使得离群检测的准确率能够在较大的范围内不会随着参数的改变而产生很大的变化, 且可以检测出数据集中的全局和局部离群点。但是 INS 算法对参数的不敏感性牺牲了一部分离群检测的准确率, 即 INS 的离群检测结果趋于稳定时检测准确率往往低于邻域参数选择合理时的其他算法。而且 INS 算法很难同时检测出局部离群点和全局离群点, 检测局部离群点时需要调整不稳定因子的设定。

通过以上分析可以得知, 现有的离群检测算法各自有各自的优势。但是, 无论是基于距离的还是基于密度的算法都存在一个参数 k 值的设置

问题，那就是如何选择一个合适的邻居个数 k 值。为了解决这一问题，本文选择结合自然邻居的思想提出适用于离群检测的普适性邻域参数选择方法。基于自然邻居形成过程无参的特性，本文提出了一种离群检测算法，该算法能够在已知离群点参数 n 的情况下无需邻域参数 k 找到数据集中的离群点。最后，算法探讨了完全无参数化的离群检测算法，在挖掘出正确的离群点的同时去除邻域参数 k 和离群点数量参数 n。

2　基于自然邻居思想的离群检测算法

2.1　自然邻居概述

自然邻居思想是笔者及课题组成员提出的一种无尺度的概念，与传统的 KNN 方法相比，该思想能够在无需邻域参数 k 的情况下构建出合理的邻居关系，为后续的数据挖掘方法提供分析基础[16]。该思想包含以下几个核心概念。

定义 1　搜索稳定状态 (search stable state)

$$(\forall x_i)(\exists x_j)(r \in N) \wedge (x_i \neq x_j) \rightarrow (x_i \in \mathrm{KNN}_r(x_j)) \wedge (x_j \in \mathrm{KNN}_r(x_i))$$

定义 2　自然邻居 (natural neighbor)

$$x_i \in NN(x_j) \Leftrightarrow (x_i \in \mathrm{KNN}_\lambda(x_j)) \wedge (x_j \in \mathrm{KNN}_\lambda(x_i))$$

定义 3　自然邻居特征值 (natural neighbor eigenvalue)

$$\lambda \triangleq r_{r \in N}\{r|(\forall x_i)(\exists x_j)(r \in N) \wedge (x_i \neq x_j) \rightarrow (x_i \in \mathrm{KNN}_r(x_j)) \wedge (x_j \in \mathrm{KNN}_r(x_i))\}$$

定义 4　自然邻居邻域图 (natural neighborhood graph)

$$e(v_i, v_j) \in E \Leftrightarrow x_j \in \mathrm{NN}(x_i)$$

在定义 1~4 中，数据集 $X=\{x_1, x_2, \cdots, x_n\}$，$\mathrm{KNN}_r(x_i)$ 代表点 x_i 的 r 邻域，即前 r 个邻居构成的集合，λ 是自然邻居特征值，$\mathrm{NN}(x_i)$ 代表点 x_i 的自然领域。自然邻居概念的定义在提出时就对该概念的扩展性进行了展望分析，而本文正是在此基本概念基础上，提出并构造加权自然邻居邻域图，并以此为基础完成无参数的离群点的检测算法。

2.2　基于加权自然邻居邻域图的离群检测算法

定义 5　加权自然邻居邻域图 (weighted natural neighborhood graph)。加权自然邻居邻域图反映了自然稳定状态时数据集中数据点之间的邻居关系，每一条加权边反映了对应的两个数据点首次互为邻居关系时的最近邻居搜索状态。加权自然邻居邻域图权值的形式化描述为

$$w(x_i, x_j) = \min(r)\{r|(x_i \in \mathrm{KNN}_r(x_j)) \wedge (x_j \in \mathrm{KNN}_r(x_i))\}$$

权值的取值范围为 $[1, \lambda]$。

定义 6　自然邻居离群因子 (natural neighbor outlier factor)。数据点 p 的自然邻居离群因子 $f(p)$ 满足：

$$f(p) = \frac{\max\limits_{q \in Q} w(p,q) \min\limits_{q \in Q} w(p,q)}{\mathrm{dgree}(p)}$$

其中，集合 $Q=\{q|e=(p,q) \in E\}$，即 Q 为数据点 p 的所有自然邻居所构成的数据子集。

在完善了加权自然邻居邻域图和自然邻居离群因子的定义后，本文提出基于加权自然邻居邻域图的离群检测算法，算法流程如图 1 所示。

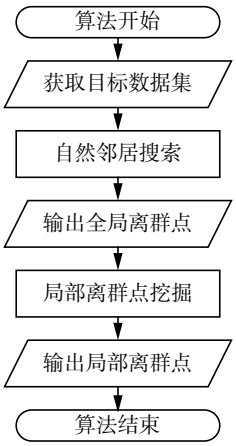


图 1　基于加权自然邻居邻域图的离群检测算法流程图
Fig. 1　Flowchart of the WNaNG outlier detection algorithm

算法 1　基于加权自然邻居领域图的离群检测

输入　目标数据集 X，离群点总数 n；

输出　加权自然邻居邻域图 G，局部离群点个数 n_l，全局离群点。

1) 初始化 k=0，并创建数据集 X 对应的 k-d 树 T；

2) 令 k=k+1，并利用 k-d 树 T 找到数据集中每个点的 k 最近邻居；

3) 分析当前的邻居关系，将互为 k 最近邻居的两点构成一条边，并记录当前的 k 作为该边的权值；

4) 重复执行步骤 2)~3)，直到数据集 X 中所有点都至少具有一个邻居，或连续 r 次未增加新的边，其中 $r=\sqrt{k-r}$；

5) 将加权自然邻居邻域图中没有边的点标记为全局离群点，并计算出剩余的局部离群点个数 n_l；

6) 根据当前所有已知边的信息构造加权自然邻居邻域图。

自然邻居搜索算法通过自然邻居搜索过程和自然邻居邻域图的简单分析，找到了全局离群点，同时给出了剩余数据的加权自然邻居邻域图。算法得到的加权自然邻居邻域图将会被用于局部离群点挖掘过程，找到数据集中最终的局部离群点。

算法 2　局部离群点挖掘算法

输入　加权自然邻居邻域图 $G=(V, E)$, 局部离群点个数 n_l;

输出　局部离群点。

1) 对邻域图进行遍历, 找到每个点的所有加权边;

2) 根据定义 6 计算所有点的自然离群因子 f;

3) 对所有点的自然离群因子进行降序排序, 前 n_l 个点即为局部离群点。

局部离群点挖掘算法首先计算加权自然邻居邻域图中点的自然邻居离群因子, 其次依照离群因子的大小得到局部离群点。

2.3　自然邻居离群因子离散图分析

在上述算法的局部离群点挖掘算法中, 剩余离群点依然需要人为设置, 如算法中则是用离群点数量参数 n 减去已经找到的全局离群点个数。为了进一步完成无参数的离群点检测算法, 本文尝试通过对自然邻居离群因子的离散图进行分析, 去除离群点数量参数 n, 使得本文提出的离群检测算法具有更强的自适应性。因此, 本文有针对性地构造了一个具有多个局部离群点的人工数据集, 并针对其局部离群点的情况和自然邻居离群因子的值进行分析。

在图 2 中可以看到, 数据点 1、2、3、4 和 162、163、164、165 为局部离群点。从图 3 中可以看出, 图 2 中提到的离群点所对应的局部离群因子均远远高于普通数据点的局部离群因子, 因此如果以自然邻居离群因子离散图作为辅助决策, 可以直观地通过其数据分布确定自然邻居离群因子的阈值, 并以此阈值作为局部离群点的判定标准, 达到去除离群点数量参数 n 的目的。这种结合图形化展示的离群点检测方法使得离群点的度量和检测能够进行更为直观地展示。

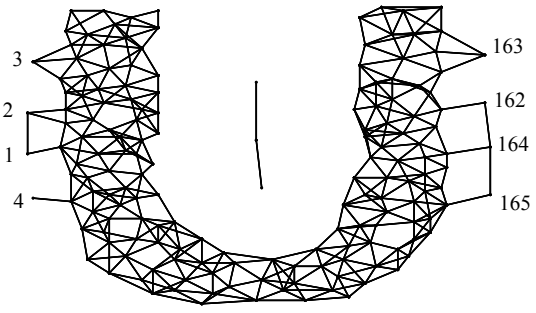


图 2　自然邻居邻域图和局部离群点示意

Fig. 2　Natural neighbor graph and local outliers

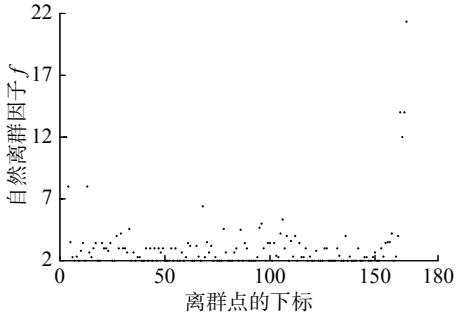


图 3　数据集中各个数据点对应的自然邻居离群因子

Fig. 3　Natural neighbor outlier factor of each data point

另外, 若希望离群检测摆脱人为的参数设置和图像化辅助决策, 在此选择对自然邻居离群因子进行降序排序, 离群度的分布形态与长尾分布具有极高的相似度, 因此可以尝试利用长尾分布的相关理论进行自然邻居离群因子的阈值确定, 实现无需人为干涉、无需邻域参数 k 和离群点数量参数 n 的自适应离群检测算法。

2.4　算法复杂度分析

本算法的总体时间复杂度为 $O(N\lg N)$。其中自然邻居搜索算法的时间复杂度为 $O(N\lg N)$, 局部离群点挖掘算法的时间复杂度为 $O(N)$。

首先, 自然邻居搜索算法在自然邻居查找和自然稳定状态的获取阶段时间复杂度为 $O(N\lg N)$。其次, 全局集群点和离群簇的挖掘均可以在邻居搜索的过程中完成, 且并不会在数量级级别增加算法的时间复杂度。因此当前阶段的时间复杂度为 $O(N\lg N)$。

在算法的第二阶段, 局部离群点挖掘算法的时间复杂度要低于上一阶段。首先, 自然邻居邻域图中的任意点最多具有 λ 条边, 最少具有一条边。因此, 自然邻居离群因子阶段的时间复杂度为 $O(\lambda N)$, 其中 λ 为一个较小的整数, 因此该阶段实际复杂度为 $O(N)$。之后的聚类分析具体操作时只需要对上一阶段中的数据簇结果进行简单的修改, 而该修改操作的时间复杂度为 $O(N)$。

3　实验结果

3.1　人工数据集实验

人工数据集的实验分为两部分。1) 着重于展示本算法自适应性的优势, 即与其他算法选择多个参数时能够获得的最好检测结果相比, 本算法能够获得与之相似甚至于更好的结果, 且无需选择邻域参数。在这一部分中, 同一数据集中多个算法均选取相同的离群点数量 n。2) 分讨论算法在无需离群点数量参数 n 的情况下利用自然邻居

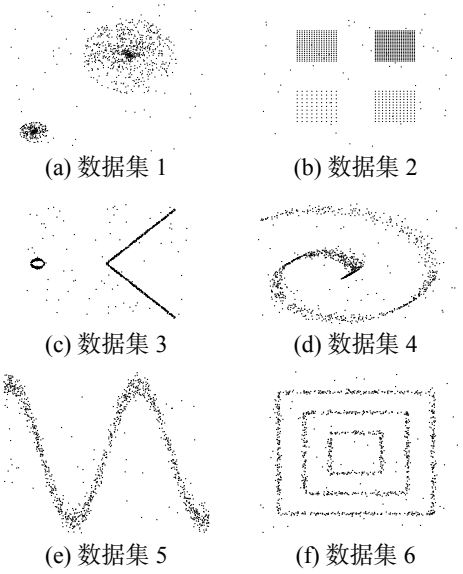
离群因子查找局部离群点的算法结果。

首先进行基于邻域参数 k 的实验展示。本实验用到的人工数据集如图 4 所示。

图 4 中的 6 个数据集均包含了多个肉眼可见的离群点，其中既有局部离群点，又有全局离群点，且这 6 个数据集具有不同的数据分布特性，因而能更准确地反映出邻域参数对算法的影响，以及本算法的自适应性在面对数据集的多样性时的实际表现。

本文将基于加权自然邻居邻域图的数据挖掘算法与 4 种离群检测算法相对比，检验本算法与当前离群检测算法之间的性能差异，被选取的对比算法为 KNN、LOF、INFLO 和 INS。

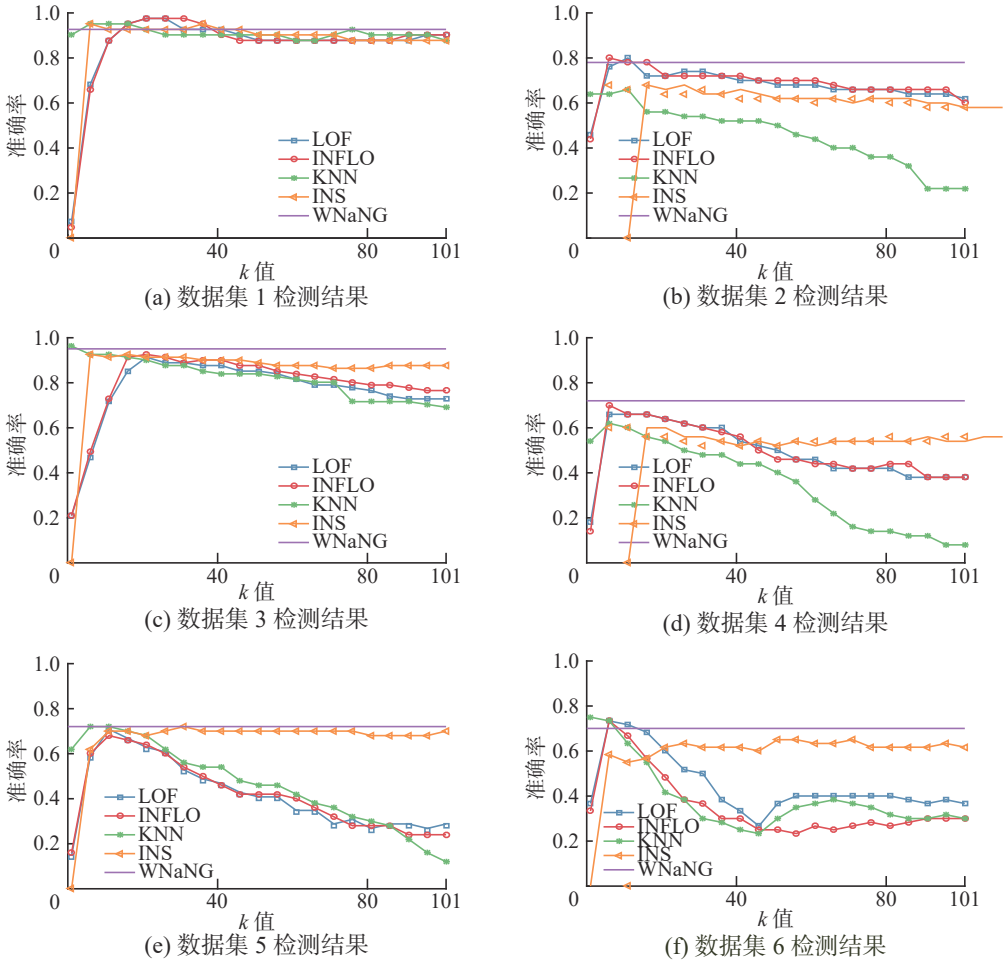
图 5 用曲线图展示了 5 种不同的离群检测算法在 6 个数据集中的检测精确率。为了更好地反映准确率随着邻域参数取值的变化而上下波动的详细情况，这里将 x 轴设定为邻域参数 k 的不同取值，y 轴则是各个算法在选取每一个 k 值时对应的离群检测准确率。本算法克服了传统的邻域选择问题，即无需在算法中设置参数 k，则在图 5 的所有子图中，WNaNG 算法的准确率是确定的，即对固定的数据集，WNaNG 算法只有一个确定的离群检测结果。为了算法对比效果的展示，WNaNG 算法的准确率采用直线进行标示。



(a) 数据集 1　(b) 数据集 2　(c) 数据集 3　(d) 数据集 4　(e) 数据集 5　(f) 数据集 6

图 4　包含离群点的 6 个人工数据集

Fig. 4　Six synthetic data sets of the outliers



(a) 数据集 1 检测结果　(b) 数据集 2 检测结果

(c) 数据集 3 检测结果　(d) 数据集 4 检测结果

(e) 数据集 5 检测结果　(f) 数据集 6 检测结果

图 5　5 种离群检测算法在取不同 k 值时离群检测准确率对比

Fig. 5　Outlier detection accuracies of five detection methods over a range of k values

纵观所有子图，WNaNG 算法的普适性高于其余算法，能够在 6 个数据集中均取得令人满意的结果。若对每一个算法在各个数据集中均选择一个最优的参数 k 与本算法相比较，在数据集 1 中其余 5 种算法的最优算法略高于本算法，而在其余几个数据集中，其最优值基本仅能与本算法取得相似的检测效果，大部分的邻域参数值所对应的检测效果均与本算法有一定的差距。

从算法对邻域参数的适应性上看，本算法完全摆脱了邻域参数的影响，并取得了令人满意的结果；INS 算法对邻域参数的敏感度较低，因此在各个数据集中，其检测结果不易随着邻域参数的变化产生剧烈波动；其余算法则对邻域参数较为敏感，特别是当数据集分布不规则时，如最后两个数据集中，邻域参数的选取会严重影响其算法结果。

产生上述情况的原因：为了降低邻域参数对算法的影响，增强算法的适应性，往往会在某些情况牺牲一部分离群检测准确率，如 INS 算法。而本文提出的 WNaNG 算法合理地利用了自适应的邻居特性，因而在保证了检测准确率的情况下移除了邻域参数的影响。本算法在与其余算法进行比较时检测结果呈现一条直线，并不是代表算法的检测结果不会随着 k 值的变化而产生变化，而是算法无需人为设置参数 k。因此可以得到结论：WNaNG 算法不仅解决了邻域参数的选取问题，更能在具有不同特性的数据集中取得稳定且准确的离群检测结果。本文将进一步讨论 WNaNG 算法利用自然邻居离群因子查找局部离群点的实验结果。

图 6 展示了 5 个不同分布特点的数据集利用自然邻居离群因子进行局部离群点检测的检测结果。从实验结果中可以看到，在前两个数据集中，自然邻居离群因子的分布相对较为均匀，对应的数据集中局部离群点的特征也较弱；而在后几个数据集中，自然邻居离群因子的分布呈现较大的差异，能够通过分布图明显地划分出一个或者多个自然邻居离群因子突变界限，而这种情况也与实际数据分布相吻合，其数据集中的局部离群点在分布上也呈现出多层级的特点，即不同范围的离群点其局部离群特征也具有较大的差异性。

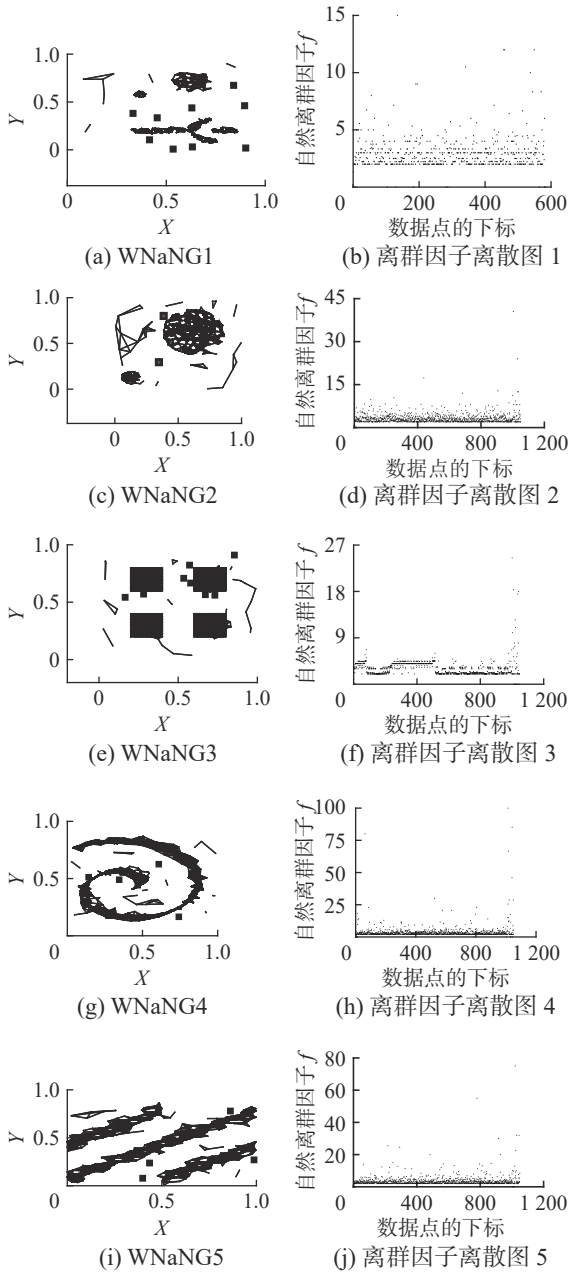


图 6　局部离群点与自然邻居离群因子离散图

Fig. 6　Local outliers and NaNOF scatter

3.2　真实数据集实验

真实数据集中本文采用的对比算法为 KNN、LOF、INFLO 和 INS 算法，算法对应的两个数据集分别为 UCI 网站的 CANCER 和 IRIS，采用的评价指标是离群检测的 ROC 曲线 (receiver operating characteristic curve)，其横、纵坐标为离群点检测率和离群点数目，通过积分面积验证数据集在对应的 k 值选择中离群检测的效率。在两个人工数据集的实验结果中，基于加权自然邻居邻域图的数据挖掘算法由于不需要邻域参数，因此图中所展示的 k 为算法自适应得出的对应数据集的自然邻居特征值，其余算法则是其对应的邻域参数 k 的

取值。

图 7 中的实验结果分为顶部和底部两组，顶部的实验结果为算法在对应的邻域参数 k 值的范围中选取得到的最差实验结果，而底部为该范围中最好的实验结果。从当前数据集可以看到，以离群点检测命中率作为检测结果时，INS 和本算法的表现相对较差。这主要是由于，在 CANCER 数据集中，部分离群点被算法归为了簇，继而难以被检测出来，而其他 3 个算法能够更快地随着参数 n 的增加而找到那部分离群点。

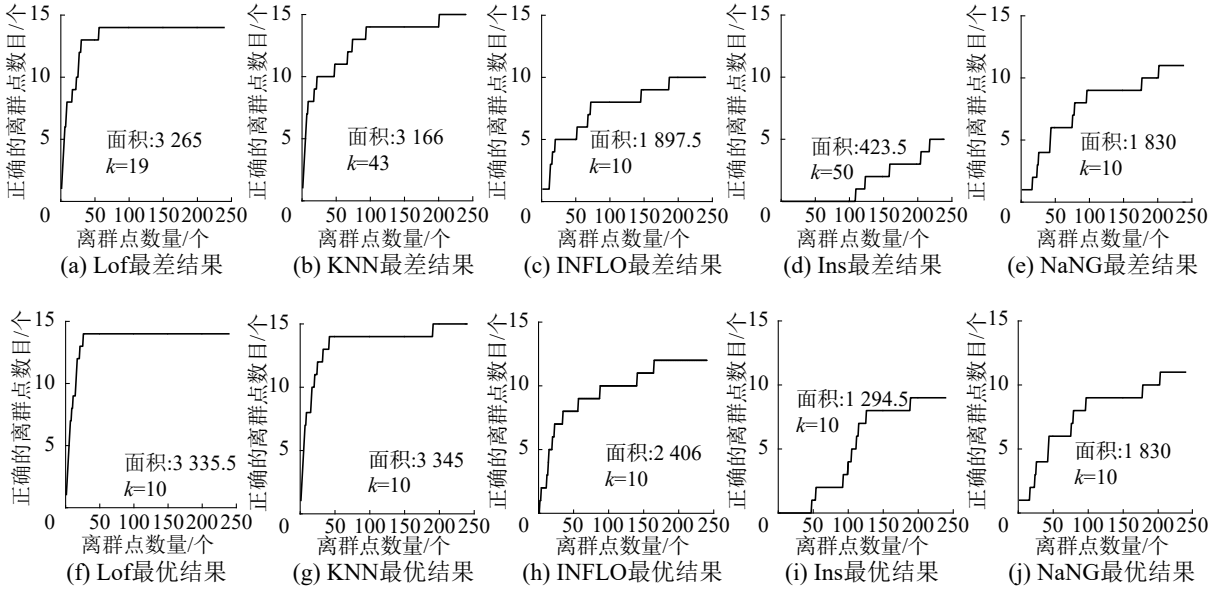


图 7　CANCER 数据集的 ROC 曲线下面积

Fig. 7　Area under the ROC curves of CANCER

图 8 的布局和图 7 相同，也是由最差实验结果和最佳实验结果组成。在 IRIS 数据集中，因为离群点中离群簇的情况比 CANCER 更少，因此 WNaNG 算法的结果有了明显的好转。特别是当 n=50 时，仅本算法就能够找到所有的离群点。

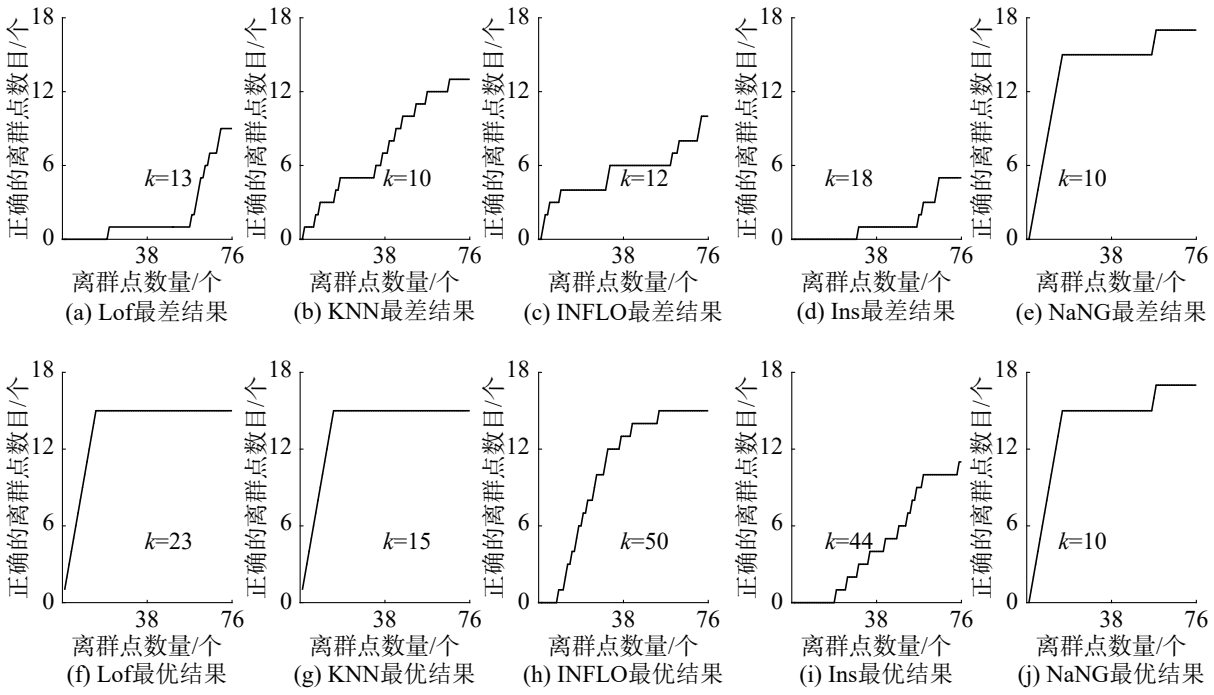


图 8　IRIS 数据集的 ROC 曲线下面积

Fig. 8　Area under the ROC curves of IRIS

总结上述两个人工数据集的实验结果可以发现: WNaNG 算法的离群检测结果不需要邻域参数, 因此不存在邻域选择影响算法效率的问题; 算法在两个数据集中均表现较为稳定, 对不同的数据集均能获得较好的效果。INS 算法需要邻域参数, 虽然其对参数的容忍度较高, 但从本实验中依然可以看到参数取值的最差情况和最好情况所对应的检测结果差距较大。其余 3 个算法在不同数据集、不同参数的情况下表现出了较大的波动, 且针对不同数据集参数的最优取值之间没有规律, 需要根据具体问题独立尝试。

4　结束语

针对离群检测中邻域参数、离群点总数参数以及局部离群点等问题, 本文结合自然邻居思想提出了一种自适应的离群检测算法 WNaNG。该算法在不同的数据集中运行时无需人为设置邻域参数, 并能够根据数据集自身的分布特征获得令人满意的检测结果。另外, WNaNG 能够更为准确地挖掘出局部离群点和全局离群点并予以区分, 这也为离群点解释、释义空间的构建等数据挖掘的后续步骤提供了强有力的支持。

参考文献:

[1] BOLTON R J, HAND D J. Statistical fraud detection: a review[J]. Statistical science, 2002, 17(3): 235–255.

[2] DENG Hongmei, XU R. Model selection for anomaly detection in wireless ad hoc networks[C]//Proceedings of 2007 IEEE Symposium on Computational Intelligence and Data Mining. Honolulu, USA, 2007: 540–546.

[3] DURAN O, PETROU M. A Time-efficient method for anomaly detection in hyperspectral images[J]. IEEE Transactions on geoscience and remote sensing, 2007, 45(12): 3894–3904.

[4] PODGORELEC V, HERICKO M, ROZMAN I. Improving mining of medical data by outliers prediction[C]//Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems. Dublin, Ireland, 2005: 91–96.

[5] NASI J, SORSA A, LEIVISKA K. Sensor validation and outlier detection using fuzzy limits[C]//Proceedings of the 44th IEEE Conference on Decision and Control. Seville, Spain, 2005: 7828–7833.

[6] KIM S, CHO N W, KANG B, et al. Fast outlier detection for very large log data[J]. Expert systems with applications, 2011, 38(8): 9587–9596.

[7] CAMPELLO R J G B, MOULAVI D, ZIMEK A, et al. Hierarchical density estimates for data clustering, visualization, and outlier detection[J]. ACM transactions on knowledge discovery from data, 2015, 10(1): 5.

[8] 苟和平, 景永霞, 冯百明, 等. 基于 DBSCAN 聚类的改进 KNN 文本分类算法 [J]. 科学技术与工程, 2013, 13(1): 219–222.
GOU Heping, JING Yongxia, FENG Baiming, et al. An improved KNN text categorization algorithm based on DBSCAN[J]. Science technology and engineering, 2013, 13(1): 219–222.

[9] 周芳芳, 高飞, 刘勇刚, 等. 基于密度-距离图的交互式体数据分类方法 [J]. 软件学报, 2016, 27(5): 1061–1073.
ZHOU Fangfang, GAO Fei, LIU Yonggang, et al. Interactive volume data classification based on density-distance graph[J]. Journal of software, 2016, 27(5): 1061–1073.

[10] 周国兵, 吴建鑫, 周嵩. 一种基于近邻表示的聚类方法 [J]. 软件学报, 2015, 26(11): 2847–2855.
ZHOU Guobing, WU Jianxin, ZHOU Song. Clustering method based on nearest neighbors representation[J]. Journal of software, 2015, 26(11): 2847–2855.

[11] 王习特, 申德荣, 白梅, 等. BOD: 一种高效的分布式离群点检测算法 [J]. 计算机学报, 2016, 39(1): 36–51.
WANG Xite, SHEN Derong, BAI Mei, et al. BOD: an efficient algorithm for distributed outlier detection[J]. Chinese journal of computers, 2016, 39(1): 36–51.

[12] 陆海青, 葛洪伟. 自适应灰度加权的鲁棒模糊 *C* 均值图像分割 [J]. 智能系统学报, 2018, 13(4): 584–593.
LU Haiqing, GE Hongwei. Adaptive gray-weighted robust fuzzy *C*-means algorithm for image segmentation[J]. CAAI transactions on intelligent systems, 2018, 13(4): 584–593.

[13] 赵冠哲, 齐建鹏, 于彦伟, 等. 移动社交网络异常签到在线检测算法 [J]. 智能系统学报, 2017, 12(5): 752–759.
ZHAO Guanzhe, QI Jianpeng, YU Yanwei, et al. Online check-in outlier detection method in mobile social networks[J]. CAAI transactions on intelligent systems, 2017, 12(5): 752–759.

[14] 张美琴, 白亮, 王俊斌. 基于加权聚类集成的标签传播算法 [J]. 智能系统学报, 2018, 13(6): 994–998.
ZHANG Meiqin, BAI Liang, WANG Junbin. Label propagation algorithm based on weighted clustering ensemble[J]. CAAI transactions on intelligent systems, 2018, 13(6): 994–998.

[15] HA J, SEOK S, LEE J S. Robust outlier detection using the instability factor[J]. Knowledge-based systems, 2014,

63(2): 15–23.

[16] 冯骥, 张程, 朱庆生. 一种具有动态邻域特点的自适应最近邻居算法 [J]. 计算机科学, 2017, 44(12): 194–201. FENG Ji, ZHANG Cheng, ZHU Qingsheng. Adaptive nearest neighbor algorithm with dynamic neighborhood [J]. Computer science, 2017, 44(12): 194–201.

作者简介：

冯骥，男，1986 年生，讲师，博士，主要研究方向为机器学习和数据挖掘。发表学术论文 10 余篇。

冉瑞生，男，1976 年生，教授，博士，主要研究方向为模式识别、机器学习。发表学术论文 20 余篇。

魏延，男，1970 年生，教授，博士，中国大数据应用联盟人工智能专家委员会委员，中国计算机学会教育专委会委员，全国高等学校计算机教育研究会理事，主要研究方向为机器学习与智能计算、数据挖掘、支持向量机理论与算法应用。主持或参与重庆市科研项目 9 项。发表学术论文 40 余篇。

2019 第九届中国智能产业高峰论坛

驱动未来，智能无界。由中国人工智能学会主办的“2019 第九届中国智能产业高峰论坛”将于 2019 年 10 月 26—27 日在西安隆重召开。

习近平总书记曾在讲话中指出：“当前，以互联网、大数据、人工智能等为代表的现代信息技术日新月异，新一轮科技革命和产业变革蓬勃推进，智能产业快速发展，对经济发展、社会进步、全球治理等方面产生重大而深远影响。”基于此，本届高峰论坛将发挥往届优势，就智能产业发展的关键问题，展开观点交锋和学术交流，推动智能产业规模化发展与智能技术的突破，加强智能科技的普及与人才培养。

随着 AI+5G 应用场景的亮相，人工智能扎根各个行业，渗透生活点滴。酷炫的智能产品、全新的交互体验，一幅 AI 生活的新画卷在慢慢铺展，人工智能已成为真正拉动经济发展的重要引擎。本届峰会，高规格、强阵容的嘉宾团将紧扣当下热点，一一勾勒人工智能的发展蓝图。

作为 2011 年学会创建的首批品牌活动之一，中国智能产业高峰论坛从纯学术活动完成了向产业应用的转型，并取得了不俗反响。峰会的召开，对我国人工智能的科学研究及在各行业落地有着指导作用与战略意义。未来，学会将继续深入学习贯彻习近平总书记重要讲话精神，将加快数字产业化、产业数字化为己任，推动数字经济和实体经济深度融合。