信阳师范学院学报(自然科学版)
JOURNAL OF XINYANG NORMAL UNIVERSITY (Natural Science Edition)

# 基于逆*K*最近邻的密度峰值异常检测方法

刘旋， 马文鹏， 杨雨晴

引用本文:
刘旋, 马文鹏, 杨雨晴. 基于逆*K*最近邻的密度峰值异常检测方法[J]. 信阳师范学院学报自然科学版, 2021, 34(2): 308-315. doi: 10.3969/j.issn.1003-0972.2021.02.023
LIU Xuan, MA Wenpeng, YANG Yuqing. Density Peak Anomaly Detection Method Based on Rknn[J]. Journal of Xinyang Normal University (Natural Science Edition), 2021, 34(2): 308-315. doi: 10.3969/j.issn.1003-0972.2021.02.023

在线阅读 View online: https://doi.org/10.3969/j.issn.1003-0972.2021.02.023

DOI:10.3969/j.issn.1003-0972.2021.02.023　　文章编号:1003-0972(2021)02-0308-08

# 基于逆 $K$ 最近邻的密度峰值异常检测方法

刘　旋[1a*],马文鹏[2],杨雨晴[1b]

(1.信阳农林学院 a. 信息工程学院; b. 财经学院,河南 信阳 464000;
2.信阳师范学院 计算机与信息技术学院,河南 信阳 464000)

**摘　要**:为提升异常检测算法在处理局部异常、异常簇和复杂分布数据集时的检测精度,降低对数据先验信息的依赖性,提出一种基于逆 $K$ 最近邻的密度峰值异常检测方法(Rknn-DP).首先结合逆 $K$ 最近邻(Rknn)改进密度峰值算法中局部密度和相对距离的计算方式,通过引入邻域信息更准确地刻画异常点的特征,然后根据特征分布选取局部密度低、相对距离高的点作为粗选异常点集合,最后通过逆 $K$ 最近邻计算粗选集合的异常因子,根据异常程度进行剪枝,排除噪声点、降低连带错误效应,自适应得到最终的异常点集.通过与 ABOD、LSCP、HBOS、IForest 等算法在真实数据集与人工数据集上的对比实验,证明了 Rknn-DP 算法的自适应性和有效性.

**关键词**:异常检测;密度峰值;逆 $K$ 最近邻(Rknn);自适应

**中图分类号**:TP391　　**文献标识码**:A　　**开放科学(资源服务)标识码(OSID)**:

# Density Peak Anomaly Detection Method Based on Rknn

**LIU Xuan[1a*], MA Wenpeng[2], YANG Yuqing[1b]**

(1a. School of Information Engineering; b. School of Finance and Economics,
Xinyang Agriculture and Forestry University, Xinyang 464000, China;
2. College of Computer and Information Technology, Xinyang Normal University, Xinyang 464000, China)

**Abstract**: In order to improve the detection accuracy of anomaly detection algorithms and reduce the dependence on data prior information when dealing with local anomalies, anomaly clusters and complex distribution data sets, a density peak anomaly detection method Rknn-DP based on inverse $K$ nearest neighbors is proposed. First of all, the algorithm improved the calculation of local density and relative distance in the density peak algorithm by Rknn, to make it more accurately describe the characteristics of the abnormal points. After that, select points with low local density and high relative distance by adaptive threshold, as rough set of abnormal points. Finally, the Rknn method is used to prune the rough selection set, eliminate the noise point, reduce the associated error effect, and adaptively get the final abnormal set. Compared with ABOD, LSCP, HBOS, IForest algorithms in real data sets and artificial data sets, the results show that the Rknn-DP, algorithm performs with higher detection and adaptability.

**Key words**: outlier detection; peak density; Rknn; adaptability

## 0　引言

异常检测是数据挖掘与知识发现研究中的重点领域之一,旨在消除数据样本中存在的噪音或者挖掘出潜在的、有意义的知识[1].异常检测方法分为有监督与无监督两类,而无监督异常检测方法因其简单、高效的特点,被广泛用于处理量大、复杂和缺乏标签的数据中.BREUING 等[2]提出了局部离群因子算法(LOF),基于可达距离、可达密度定义局部离群因子,进而度量样本的异常程度,实现异常检测.CHONG 等[3]提出了一种将稀疏表示与随机游走相结合的异常检测方法,利用稀疏线性组合得到数据之间的非对称亲和矩阵,然后构造加权有向图,采用随机游走检测异常值.DU 等[4]提出了

**收稿日期**:2019-11-21;**修订日期**:2020-10-26;*.**通信联系人**,E-mail: 31749325@qq.com

**基金项目**:国家自然科学基金项目(31660239,61702438);河南省教育厅人文社会科学研究一般项目(2020-ZDJH-353)

**作者简介**:刘旋(1991—),男,河南信阳人,讲师,硕士,主要从事多粒度计算、模式识别等研究.

一种利用统计参数进行局部异常检测的方法,结合聚类和密度峰值方法实现了大数据中的异常检测. ZHAO 等[5]提出了一种基于局部选择性集成学习的异常检测算法,根据随机特征子空间的最近邻定义样本的局部区域,然后选择区域内最优的基学习器检测结果进行输出.涂晓敏等[6]提出了一种基于方形邻域和裁剪因子的异常检测方法,通过二次筛选快速准确地得到结果.

针对复杂数据集中存在的多类聚、非球状分布、无标签等特点,传统的无监督异常检测方法大多需要根据异常数据比例人工设置阈值,使得检测结果的优劣很大程度上依赖于数据分布的先验信息.密度峰值聚类(Density Peaks Clustering, DPC)[7]是一种基于密度的聚类算法,具有简单高效、无须指定参数、有效识别任意分布等优点,能够实现聚类中心的自动选取. DPC 采用密度与距离选择类中心时,也能刻画异常样本的特征.

综上,本文在 DPC 的基础上,提出了一种基于逆 $K$ 最近邻的密度峰值异常检测方法 Rknn-DP (Rknn Density Peaks). Rknn-DP 主要贡献包括:1)引入逆 $K$ 最近邻改进 DPC 算法中局部密度与相对距离,以此作为特征共同度量样本的异常性,避免截断距离选取对密度的影响,而且结合邻域信息扩大异常与正常样本之间的特征差异;2)根据异常特征分布,自适应设置阈值对数据集进行异常样本的粗选,并结合逆 $K$ 最近邻对粗选结果进行剪枝,排除噪声干扰,得到优化后的精选结果.

## 1 DPC 算法分析

密度峰值聚类算法认为每个类都包含一个最大的密度点作为类中心,每个类中心都吸引并连接其周围密度较低的点,且不同的类中心点都相对较远.

### 1.1 基本定义

DPC 定义类中心需满足 2 个条件:(1)类中心点的密度大于周围邻居点的密度;(2)类中心点与更高密度点之间的距离相对较大.因此,采用局部密度和相对距离能更准确地刻画中心点的特征.假设数据集为 $S$,数据点 $i$ 的局部密度 $\rho_i$ 定义如下:

$$\rho_i=\sum_{j\in S\setminus\{i\}}\chi(d_c-d_{ij}),\qquad \chi(x)=\begin{cases}1, & x>0,\\ 0, & x<0,\end{cases} \tag{1}$$

其中:$d_{ij}$ 表示数据点 $i$ 和 $j$ 的欧式距离;$d_c$ 代表截断距离,$d_c$ 通常取值为全体数据点之间的相互距离升序后的前 2%;$\chi$ 为激活函数,所以与 $i$ 距离小于 $d_c$ 的点越多,则 $\rho_i$ 越大.数据点 $i$ 的相对距离 $\delta_i$ 的计算如下:

$$\delta_i=\begin{cases}\min\limits_{j:\rho_j>\rho_i}\{d_{ij}\}, & \rho_i\neq\max(\rho),\\ \max\limits_{j\in S\setminus\{i\}}\{d_{ij}\}, & \rho_i=\max(\rho),\end{cases} \tag{2}$$

这里:$\max(\rho)$表示 $S$ 中的最大局部密度值;$\delta_i$ 表示在所有局部密度大于 $i$ 的数据点中,与 $i$ 距离最近的点的距离.通过绘制 $S$ 中全体数据点的($\rho_i$, $\delta_i$)坐标决策图发现,当某些数据点密度和距离较大时,会被判定为聚类中心.

### 1.2 异常点定义

DP 算法利用 $\rho_i$ 和 $\delta_i$ 刻画聚类中心的同时,也发现局部密度越小、相对距离越大的数据点越偏离类中心,越有可能是异常点.图 1 描述了一个 2 维人工数据集 Test 的分布情况,其中 $V_1$ 和 $V_2$ 表示每个数据点的 2 维属性,倒三角代表异常点,空心圆代表正常点.根据密度大小和远离类中心的距离,可以将异常点分为三类:异常簇(圆形区域,局部密度较高)、显著异常点(矩形区域,远离所有类中心)、局部异常点(椭圆区域,相对某一类是异常点).

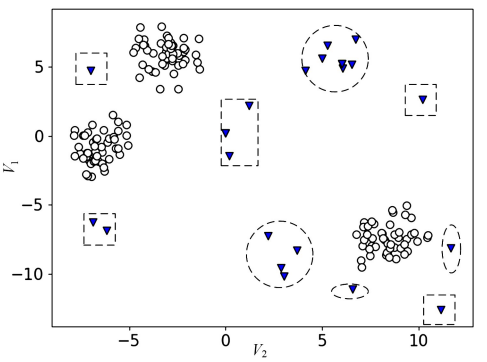



图 1 异常点分布

**Fig. 1 Distribution of outliers**

文献[8,9]采用 $\rho_i$ 和 $\delta_i$ 度量数据点的异常值,通过人工设定相应阈值选取异常点,但密度峰值特征的计算受到截断距离 $d_c$ 选取和邻域信息的影响,对异常簇、局部异常点的检测结果不理想.因此如何改进 DP 算法中密度与距离的计算方式,更准确地刻画三类异常点的特征是本文的研究重点.

## 2 Rknn-DP 算法

### 2.1 相关定义

**定义 1**(*逆 $K$ 最近邻*(Rknn)[10]) 对象 $p\in S$

的 Rknn 为 $\mathrm{Rknn}(p)=\{q\mid p\in \mathrm{knn}(q), q\neq p, q\in S\}$，$\mathrm{knn}(q)$为 $q$ 的 $K$ 最近邻.

当对象 $q$ 属于另一个对象 $p$ 的 $K$ 最近邻，那么 $p$ 就是 $q$ 的一个逆 $K$ 最近邻. 根据 $K$ 的取值，$q$ 的逆 $K$ 最近可能为零或多个.

**定义 2**　基于逆 $K$ 最近邻的局部密度(RKnn Local Density). 样本 $i$ 的局部密度定义如下：

$$\rho_i=\exp\left(-\frac{1}{k}\sum_{j\in \mathrm{Rknn}(i)} d_{ij}^2\right). \tag{3}$$

**定义 3**　基于 $K$ 近邻的相对距离(Knn Distance). 样本 $i$ 的相对距离定义如下：

$$\delta_i=\begin{cases}\dfrac{1}{k}\displaystyle\sum_{j\in \rho_{\mathrm{knn}}(i)} d_{ij}, & \mid \rho_{\mathrm{knn}}(i)\mid \geqslant k,\\ \dfrac{1}{\mid \rho_{\mathrm{knn}}(i)\mid}\displaystyle\sum_{j\in \rho_{\mathrm{knn}}(i)} d_{ij}, & 1<\mid \rho_{\mathrm{knn}}(i)\mid <k,\\ \max(d_{ij}), & \mid \rho_{\mathrm{knn}}(i)\mid =0.\end{cases} \tag{4}$$

这里：$\rho_{\mathrm{knn}}(i)$表示局部密度大于 $\rho_i$ 且距离 $i$ 最近的 $k$ 个样本集合；当 $i$ 具有最大局部密度时，$\delta_i$ 表示与 $i$ 距离最大的样本与 $i$ 之间的距离；否则，$\delta_i$ 表示局部密度大于 $i$、距离 $i$ 最近的 $k$ 个样本与 $i$ 之间的距离均值.

**定义 4**　异常因子(An).

假设异常 $i$ 的正常样本邻居集合定义为

$$\mathrm{NN}(i)=\mid \mathrm{normal}\cap \mathrm{Rknn}(i)\mid,$$

$i$ 的异常因子 $\mathrm{An}(i)$为

$$\mathrm{An}(i)=\begin{cases}\exp\left(-\displaystyle\sum_{j\in \mathrm{NN}(i)}(\rho_i/\rho_j)\right), & \mid \mathrm{NN}(i)\mid \geqslant 1,\\ 1, & \mid \mathrm{NN}(i)\mid =0,\end{cases} \tag{5}$$

其中：normal 表示正常样本集合，$\mathrm{An}(i)\in[0,1]$的值越大，表明样本 $i$ 的异常度越高. 记 anomaly 表示异常样本集合，

## 2.2　异常特征定义

(1)局部密度特征

截断距离 $d_c$ 的取值决定样本的局部密度计算精度，而原文中 $d_c$ 的选取缺乏理论依据，针对异常检测，如果 $d_c$ 过大，则样本间的局部密度差异较小，导致异常样本检测率低，如果 $d_c$ 过小，则局部密度差异较大，导致正常样本的误检率高.

为了消除 $d_c$ 对密度的影响，本文引入定义 1(逆 $K$ 最近邻)计算样本的局部密度，与 knn 不同，Rknn 的大小反映了距离对象周围较近样本的数目，所以逆 $K$ 最近邻的数目能够反映对象局部的密度且对参数 $K$ 的取值不敏感.

Test 中异常样本的 Rknn 分布如图 2，虚线代表 $K$ 近邻的区域范围，当 $K=6$ 时，显著异常样本 $q_1\sim q_3$ 的 Rknn=0，边界异常样本 $p_1$ 和 $p_2$ 的 Rknn=1，异常簇样本 $o_1\sim o_4$ 的 Rknn 分别为 2、3、4 和 6，说明了逆 $K$ 最近邻能够准确反映样本的局部密度，因此本文采用定义 2 计算样本的 $\rho$，$\rho$ 值越大表明该样本的局部密度越大.

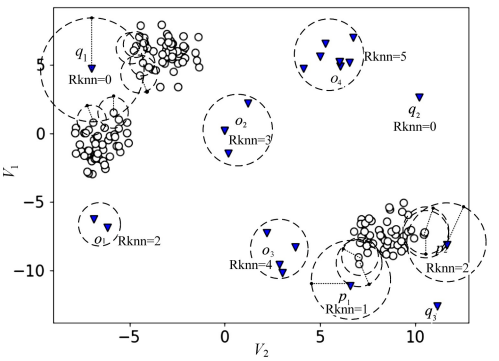



**图 2　Rknn 密度范围**

**Fig. 2　Rknn density range**

(2)相对距离特征

异常簇具有局部密度高，距离密度大的正常群体较远的特点，仅采用局部密度特征难以区分边缘正常数据与中心异常簇数据，而相对距离 $\delta$ 能够反映样本与高密度区域的远近，进而度量差异. 图 3 中，边缘正常样本 $n_1$、异常簇 $o_2$ 和 $o_4$ 的 Rknn 分别为 2、3 和 5，虽然它们的局部密度相近，但通过公式(3)计算的相对距离为 $\delta(o_2)=d_3$、$\delta(o_4)=d_2$、$\delta(n_1)=d_1$，$d_1<d_2<d_3$，进而较直观地反映异常簇与边缘正常样本的差异.

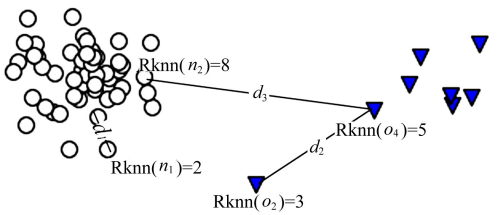
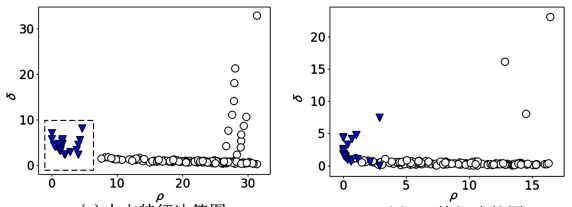



**图 3　Rknn 相对距离**

**Fig. 3　Rknn relative distance**

综上，本文采用定义 3 计算全体数据的 $K$ 近邻相对距离特征，通过结合 knn 思想改进 DPC 中相对距离的计算方式，考虑了每个样本的邻域，降低连带错误效应和异常簇高密度样本的干扰，提高异常簇和局部异常样本与正常样本之间的距离差距.

DPC 和 Rknn-DP 计算 Test 的特征分别如图 4(a)和图 4(b)所示,DPC 中部分异常与正常样本的特征出现交叠,难以区分,Rknn-DP 通过引入邻域信息改进局部密度和相对距离的计算方式,能够进一步扩大异常与正常样本的特征差距,将异常样本集中刻画在左下角的矩形区域内,证明 Rknn-DP 的密度与距离计算方式能够有效刻画三类异常的特征.




(a)本文特征决策图　(b)DPC特征决策图

**图 4　异常特征分布**

**Fig. 4　Abnormal Feature Distribution**

## 2.3　异常检测过程

为提升异常检测的效率,实现自适应选取,Rknn-DP 的检测过程为粗选和剪枝两个阶段:

(1)粗选阶段:首先计算数据集 $S$ 中各个样本的 Rknn 集合;然后通过 Rknn 计算全体样本的特征集合$(\rho,\delta)$;其次,根据特征分布设定密度阈值 $OT_\rho$、距离阈值 $OT_\delta$;最后选取局部密度小于 $OT_\rho$、相对距离大于 $OT_\delta$ 的样本作为粗选异常样本(anomaly),其余作为粗选正常样本(normal).

观察图 4(b)可知,异常样本集中分布在决策图的左下角,局部密度越小且相对距离越大的样本更可能为异常.由文献[11]得出,整体数据集的逆 $K$ 最近邻服从高斯分布,可推导出 $\rho$ 和 $\delta$ 也近似服从高斯分布,根据高斯分布的 $3\sigma$ 规则和大量实践经验,本文选取 $OT_\rho=2\sigma(\rho)$(2 倍标准差)作为密度阈值,$OT_\delta=\mu(\delta)$(均值)作为距离阈值,保证粗选异常样本的选取质量.

(2)剪枝阶段:根据粗选阶段的结果,采用定义 4 计算粗选 anomaly 中全体样本的异常因子 An,将 $An<T_{An}$(异常因子阈值)的样本从 anomaly 中删除,并归为 normal.通过循环迭代剪枝过程,完成二次筛选得到最终精选后的异常集合.

在定义 4 中,当 $NN(i)=0$ 时,表示为孤立点,远离正常群体;当 $NN(i)\geqslant 1$ 时,表示 $i$ 在正常样本邻域内,$NN(i)$ 的数量越多,说明 $i$ 的正常样本邻居越多,$\rho_i/\rho_j$ 值越大,说明与其正常邻居的局部密度越相似,越有可能属于正常样本,所以 $An(i)$ 的大小与样本的异常程度成正比,通过设置阈值 $T_{An}$ 能够排除边界噪声样本和粗选 anomaly 的误分样本,通常$|anomaly|\ll|S|$,所以仅计算粗选 anomaly 的异常因子,能够大大提升算法效率,自适应得到最终的异常检测结果.

算法　自适应密度峰值异常检测(Rknn-DP)

输入:数据集 $S=\{s_i,i=1,2,\cdots,n\}$、阈值 $T_{An}$、参数 $K$

输出:异常样本集合 $anomal=\{a_1,a_2,\cdots,a_n\}$

1. 计算 $S$ 全体样本两两之间的距离矩阵 Dist,根据 Dist 和定义 1 求得样本 $K$ 近邻和逆 $K$ 最近邻,

$Knn=\{knn(s_i),i=1,2,\cdots,n\}$,

$RKnn=\{Rknn(s_i),i=1,2,\cdots,n\}$.

2. 根据定义 2 和 3 计算全体样本的异常特征 $\rho=\{\rho_{s_i},i=1,2,\cdots,n\}$,$\delta=\{\delta_{s_i},i=1,2,\cdots,n\}$,然后计算密度阈值 $OT_\rho=2\sigma(\rho)$ 和距离阈值 $OT_\delta=\mu(\delta)$.

3. 遍历 $S$,当样本 $S_i$ 的 $\rho_{s_i}<OT_\rho$ 且 $\delta_{s_i}>OT_\delta$ 时,判定为粗选异常样本 $anomal=\{a_i,i=1,2,\cdots,M\}$,$M=M+1$,否则为正常样本 $normal=\{n_i,i=1,2,\cdots,N\}$,$N=N+1$.

4. 遍历 anomal,若 $a_i$ 满足

$An(a_i)<T_{An}$,

则从 anomaly 中删除,$M=M-1$,并重新归为 normal,$N=N+1$,循环该步骤直到前后两次迭代的 anomaly 数据量不变,输出最终 anomaly.

## 2.4　时间复杂度分析

Rknn-DP 的时间复杂度分为特征计算和异常检测两个阶段:

1. 首先计算样本之间的距离矩阵,时间复杂度为 $O(n^2)$,然后计算出样本的 Knn 邻居和 RKnn 集合,并求得 $\rho$ 和 $\delta$ 的特征集合,时间复杂度为 $O(2n^2+n+n)$,$n$ 为数据集中的样本个数;

2. 首先计算阈值 $OT_\rho$、$OT_\delta$,遍历数据集进行筛选,得到粗选异常集合,时间复杂为 $O(n)$,然后计算粗选异常样本的 RKnn 中属于正常样本的集合,求得样本的异常因子且循环 $k$ 次,复杂度为 $O(km\ln(n-m))$,$m$ 表示粗选异常集合的样本个数.综上,Rknn-DP 的时间复杂度为 $O(2n^2+n+n)+O(n)+O(km\ln(n-m))$,一般情况下,异常个数远远小于样本个数,即 $m\ll n$,而且循环次数 $k$ 也较小,所以时间复杂度可记为 $O(n^2)$.

## 3　结果与分析

为验证 Rknn-DP 的可行性,与 ABOD[12](角度方差距离)、CBLOF[13](密度聚类)、LSCP[4](集成学习)、HBOS[14](直方图统计)和 IForest[15](决策树)等异常检测算法在多种数据集下进行实验对比,其中 LSCP 算法的基类检测器为 LOF. 本文采用查全率、误检率和运行时间作为衡量异常检测算法好坏的标准,查全率(RR)表示被正确检测的异常样本数量占整个异常集合的比例;误检率(FR)表示正常样本被误检为异常样本的数量占整个正常样本集合的比例,RR 的值越高、FR 的值越低,则表明算法的检测性能越好. 实验环境:Windows10、Inter Corei5-7400、16 GB、Python、PYOD[16](算法库).

此外,为了验证 Rknn-DP 异常检测的准确性,本文所有实验均设置固定的 RKnn 的参数 $K$ 值和异常因子阈值 $T_{An}=0.2$(根据大量实践经验选取),不需要数据集的任何先验信息,而所有对比算法均根据不同数据集中的异常样本占比,人工选取最优 TOP-N[17] 作为检测结果.

### 3.1　网络日志数据集

本数据集由某高校校园卡系统中的网络日志组成,经网络管理中心授权采集,日志中记录了学生 ID、访问时间、URL 地址、流量使用、消费金额、图书馆出入时间等大量信息. 网络日志反映了学生的生活轨迹、行为习惯,本文将异常定义为与大部分学生群体在某些网络行为上差异较大的个别学生对象,这里异常本身没有褒贬之分. 本实验旨在采用异常检测方法预测学生网络日志数据中的异常行为,并及时进行预防决策.

本文随机选取 2019 年某一周内的网络日志作为实验数据,日志文件的体积达到 1.7 TB. 为了便于实验分析,将 URL 进行分类(门户网站、搜索引擎、影视音乐、学习资源、购物消费、微博博客、游戏娱乐),经过数据清洗(静态资源、噪声删除)、统计归一化等处理,得到包含 12 个特征,共计 2985 条(人)可用数据,主要特征包括:性别、URL 分类的访问次数(7 维)、上行流量、下行流量、校园卡消费金额、图书馆出入时间等. 根据学生成绩与平时表现,通过前期的专家统计分析,得出学生行为异常样本为 67 条(人).

首先,研究 Rknn 中参数 $K$ 的取值对算法性能的影响. 从图 5 可知,Rknn-DP 随着参数 $K$(占样本总量的百分比)的不断提升,RR 和 FR 在 $K=3\%$ 以后逐渐稳定,当 $K=3\%$ 时,能够得到最高的 RR(88.01)和较低的 FR(0.45),说明了参数 $K$ 对实验结果的影响较低,具有良好的参数鲁棒性,后续实验的参数 $K$ 均固定取值为 3%.

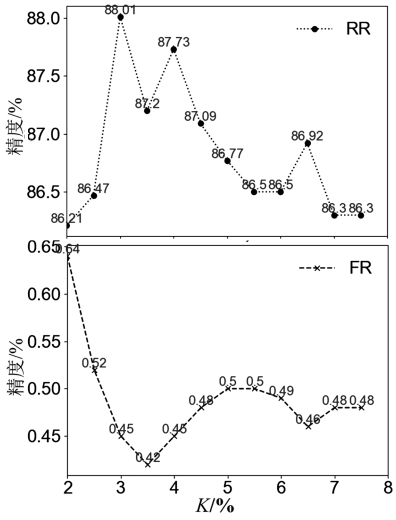



**图 5　参数 $K$ 的敏感性测试**

**Fig. 5　Sensitivity testing of parameter $K$**

然后与上述 5 种异常检测算法进行对比,测试算法的准确性,结果如表 1 所示. Rknn-DP 的查全率均高于其他异常检测算法,且误检率相对较低,证明了 Rknn-DP 能够得到理想的异常样本候选集,但 Rknn-DP 的时间复杂度相对较高,异常特征计算阶段耗时略长,导致算法整体效率较低. 整体而言,验证了 Rknn-DP 的可行性.

**表 1　异常检测方法在网络日志上的实验结果**

**Tab. 1　Results of anomaly detection on network log**

| Measure | RR/% | FR/% | Time/s |
|---|---|---|---|
| ABOD | 82.08 | 0.40 | 2.39 |
| CBLOF | 79.10 | 0.46 | 1.44 |
| LSCP | 86.56 | 0.30 | 3.31 |
| HBOS | 77.61 | 0.51 | 1.12 |
| IForest | 83.58 | 0.36 | 0.54 |
| Rknn-DP | 88.05 | 0.45 | 3.06 |

### 3.2　人工数据集

为验证 Rknn-DP 算法在处理不同分布类型数据集时的异常检测效果,本文选取了 4 种分布类型的二维人工数据集进行实验,分别用 $V_1$、$V_2$ 表示数据点的 2 维属性. 原始人工数据集分布如图 6 所

示，第一组数据共计 2 个类、258 个数据对象、15 个异常点，第二组数据共计 1 个类、147 个数据对象、20 个异常点，第三组数据共计 3 个类、1000 个数据对象、50 个异常点，第四组数据共计 3 个类、153 个数据对象、23 个异常点. 上述人工数据集中的正常点(约占 95%)用空心圆标志，异常点(约占 5%)用实心倒三角标志. 对比实验结果如图 7(a)～7(e)所示.

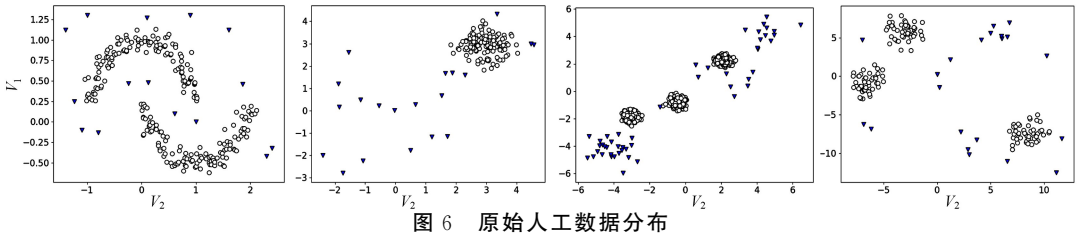


图 6　原始人工数据分布

**Fig. 6　Raw artificial data distribution**

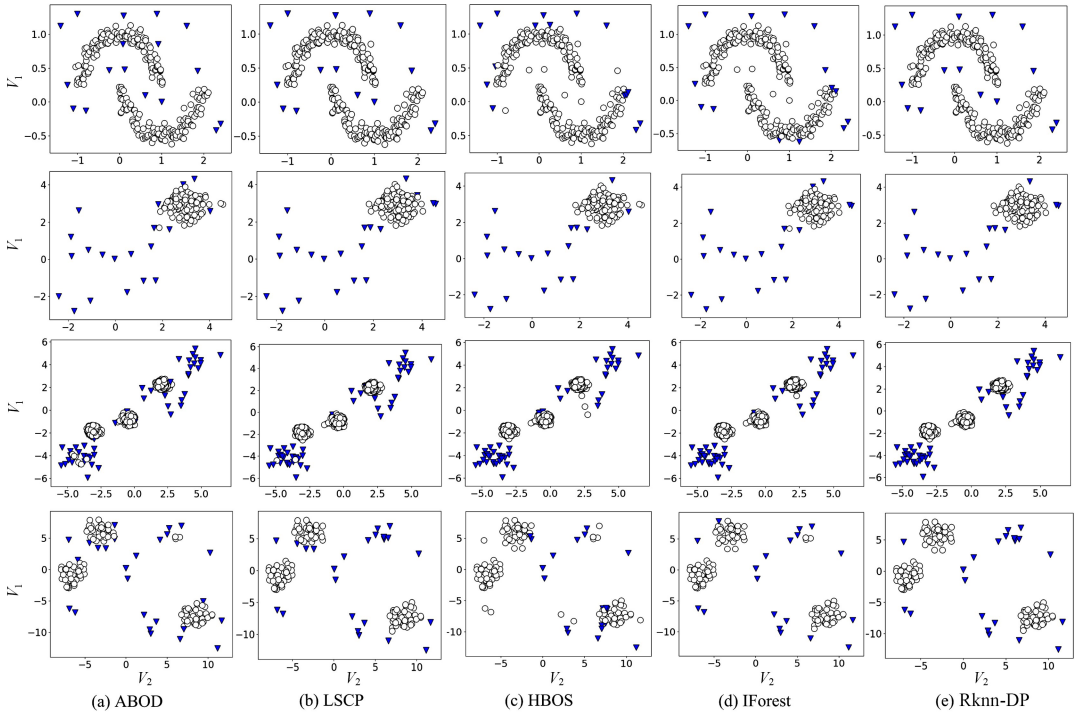



图 7　人工数据检测结果

**Fig. 7　Manual data detection results**

通过观察图 7 可知，在处理四组不同分布类型的数据集时，ABOD 算法对显著异常、局部异常的查全率较高，但对异常簇的检测结果不理想，容易受到相邻异常样本的干扰. LSCP 对所有数据集的检测结果均较好，但在对边界噪声样本的误检率略高. HBOS 算法对局部异常的检测结果较差，对显著异常和异常簇的查全率较高. IForest 算法与 HBOS 算法类似，其整体检测结果优于 HBOS. Rknn-DP 算法能有效处理多种分布数据，准确、自适应地检测出三类异常样本，且抗噪性较好，整体结果优于 LSCP 算法. 几种算法的平均检测结果为：ABOD(RR＝85. 18，FR＝1. 17)，LSCP(RR＝93. 51，FR＝0. 62)，HBOS(RR＝85. 33，FR＝1. 10)，IForest(RR＝90. 74，FR＝0. 48)，Rknn-DP(RR＝99. 07，FR＝0. 00)，证明了 Rknn-DP 处理多种分布类型数据集的有效性.

### 3. 3　公共数据集

进一步验证 Rknn-DP 在处理不同维度、异常比例数据集时的准确性和自适应性，采用 ELKI 开源数据库[18]中的真实异常数据集进行实验，所用数据集均已预处理(标准化、重复和缺失值清洗)，包括 Parkinson、Lymphograph、Ionosphere、Stamps、PageBlocks、KDD-Cup99 等 6 组数据集. 为了降低实验复杂性，对 KDD-CUP99 数据集进行

了适当调整,调整后的数据集共计 9812 个样本,其中包含 594 个网络入侵数据(异常样本),占比约 5%.本文所用 6 种数据集的具体规模如表 2 所示.实验结果如表 3 所示,最后一行 AVG 为各个算法在所有数据集上检测指标结果的平均值.

由表 2 可知,Parkinson 数据集的数据量少、维度高,导致几种算法的准确率和误检率均不理想,Rknn-DP 的误检率最低,但检测率低于 ABOD.针对异常占比低、特征差异性显著的 Lymphography 数据集,除 ABOD 外的对比算法查全率和误检率均较好,但仍低于 Rknn-DP.LSCP 在处理异常占比高(35%)的 Ionosphere 数据集时,取得最高的检测率和较低的误检率,而 Rknn-DP 虽然检测率略低于 MCD,但误检率最低.随着正常数据与异常数据特征差异性的降低,导致对比算法在处理 Stamps 和 PageBlocks 数据集时,查全率均显著降低,而 Rknn-DP 仍能保证较高的查全率.针对人工筛选的 KDDCup99 高维数据集,各个算法的检测结果均较好,LSCP 整体指标最优.综上,相较于现有的异常检测算法,Rknn-DP 的平均指标 AVG 最优,证明 Rknn-DP 能够有效、自适应地检测不同的数据规模、异常占比数据集中的异常样本.

**表 2　数据分布情况**

**Tab. 2　Distribution of data**

| 数据集 | 样本总数 | 离群样本数 | 特征维度 |
|---|---|---|---|
| Parkinson | 60 | 12 | 22 |
| Lymphography | 148 | 6 | 18 |
| Ionosphere | 351 | 126 | 32 |
| Stamps | 340 | 31 | 9 |
| PageBlocks | 5473 | 560 | 10 |
| KDDCup99 | 9812 | 594 | 38 |

**表 3　几种异常检测方法的性能比较**

**Tab. 3　Performance comparison of several anomaly detection methods**

| 数据集 | ABOD | | CBLOF | | LSCP | | HBOS | | IForest | | Rknn-DP | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RR/% | FR/% | RR/% | FR/% | RR/% | FR/% | RR/% | FR/% | RR/% | FR/% | RR/% | FR/% |
| Parkinson | 66.67 | 10.41 | 50.00 | 12.50 | 50.00 | 12.50 | 41.67 | 14.58 | 50.00 | 12.50 | 50.00 | 4.16 |
| Lymphography | 50.00 | 2.02 | 83.33 | 0.70 | 83.33 | 0.70 | 83.33 | 0.70 | 83.33 | 0.70 | 100 | 0.67 |
| Ionosphere | 84.12 | 10.22 | 62.70 | 20.08 | 87.30 | 5.77 | 45.23 | 29.33 | 66.67 | 17.33 | 84.95 | 2.56 |
| Stamps | 25.80 | 10.67 | 51.61 | 4.41 | 70.97 | 2.65 | 38.70 | 5.59 | 29.03 | 7.11 | 96.77 | 6.76 |
| PageBlocks | 33.26 | 7.51 | 41.83 | 6.11 | 63.21 | 3.76 | 35.99 | 7.26 | 42.61 | 6.57 | 73.57 | 5.13 |
| KDDCup99 | 80.47 | 1.16 | 83.33 | 0.99 | 84.16 | 0.99 | 82.82 | 1.02 | 81.81 | 1.08 | 84.05 | 0.89 |
| AVG | 56.72 | 7.00 | 62.13 | 7.47 | 73.16 | 4.40 | 54.62 | 9.75 | 58.91 | 7.55 | 81.56 | 3.36 |

## 4　结束语

本文提出的 Rknn-DP 算法通过结合逆 $K$ 最近邻和密度峰值算法来刻画异常样本的特征,将异常检测过程分为粗选和剪枝两个阶段,降低了时间复杂度,且无须设置异常样本选取比例或阈值,能够有效、自适应地发现数据集中的异常样本.实验结果表明,本文算法在网络日志数据集、人工数据集和公共数据集的实验均取得了较理想的结果,有效处理不规则形状和复杂分布的多维数据集.但本文算法对非数值属性的处理较差,如何进一步提升算法效率、处理多种数据类型以及设计不确定性异常的检测方法[19]是未来的研究重点.

**参考文献:**

[1]　陈斌, 陈松灿, 潘志松, 等. 异常检测综述[J]. 山东大学学报(工学版), 2009, 39(6): 13-23.
CHEN Bin, CHEN Songcan, PAN Zhisong, et al. Survey of outlier detection technologies[J]. Journal of Shandong University (Engineering Science), 2009, 39(6): 13-23.

[2]　BREUNIG M M, KRIEGEL H P, NG R T, et al. LOF: identifying density-based local outliers[J]. ACM SIGMOD

Record，2000，29(2)：93-104.

[3] YOU C，ROBINSON D P，VIDAL R. Provable self-representation based outlier detection in a union of subspaces [C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR)，Honolulu，HI，USA，2017：4323-4332.

[4] DU H，ZHAO S，ZHANG D. Robust local outlier detection[C] // IEEE International Conference on Data Mining Workshop (ICDMW)，Atlantic City，NJ，2015：116-123.

[5] ZHAO Y，NASRULLAH Z，HRYNIEWICKI M K，et al. LSCP：Locally selective combination in parallel outlier ensembles[C] // Proceedings of the 2019 SIAM International Conference on Data Mining，2019：585-593.

[6] 涂晓敏，石鸿雁. 基于方形邻域和裁剪因子的离群点检测方法[J]. 小型微型计算机系统，2019，40(1)：186-189.
TU Xiaomin，DAN Hongyan. Square neighborhood and pruning factor based outlier detection algorithm[J]. Journal of Chinese Computer Systems，2019，40(1)：186-189.

[7] RODRIGUEZ A，LAIO A. Clustering by fast search and find of density peaks[J]. Science，2014，344(6191)：1492-1496.

[8] ZHAO X，WANG G，LI Z. Unsupervised network anomaly detection based on abnormality weights and subspace clustering[C] // Sixth International Conference on Information Science and Technology(ICIST)，Dalian，2016：482-486.

[9] 刘凤魁，邓春宇，王晓蓉，等. 基于改进快速密度峰值聚类算法的电力大数据异常值检测[J]. 电力信息与通信技术，2017，15(6)：36-41.
LIU Fengkui，DENG Chunyu，WANG Xiaorong，et al. Outlier detection of smart grid big data based on improved fast search and find density peaks clustering algorithm [J]. Electric Power Information and Communication Technology，2017，15(6)：36-41.

[10] KORN F，MUTHUKRISHNAN S. Influence sets based on reverse nearest neighbor queries[J]. ACM SIGMOD Record，2000，29(2)：201-212.

[11] 卢建云. 基于邻域的离群检测与聚类算法研究[D]. 重庆：重庆大学，2017.
LU Jianyu. Research on outlier detection and clustering algorithm based on neighborhood [D]. Chongqing：Chongqing University，2017.

[12] KRIEGEL H P，SCHUBERT M，ZIMEK A. Angle-based outlier detection in high-dimensional data[C] // Proceedings of the 14th ACM SIGKDD International Conference on Knowledge discovery and data mining，Las Vegas，Nevada，2008：444-452.

[13] HE Z，XU X，DENG S. Discovering cluster-based local outliers[J]. Pattern Recognition Letters，2003，24(9/10)：1641-1650.

[14] GOLDSTEIN M，DENGEL A. Histogram-based outlier score(hbos)：A fast unsupervised anomaly detection algorithm[C]// KI—2012：Poster and Demo Track，2012：59-63.

[15] LIU F T，TING K M，ZHOU Z H. Isolation-Based anomaly detection[J]. ACM Transactions on Knowledge Discovery From Data，2012，6(1)：1-39.

[16] ZHAO Y，NASRULLAH Z，PYOD L Z. A python toolbox for scalable outlier detection[J]. Journal of Machine Learning Research，2019，20(96)：1-7.

[17] YAN Y，CAO L，RUNDENSTEINER E A. Scalable top-n local outlier detection[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining，Halifax，NS，Canada，2017：1235-1244.

[18] ELKI Data Mining. Outlier detection data sets [EB/OL]. [2019-03-28]. http://www.dbs.ifi.lmu.de/research/outlier-evaluation/DAMI/.

[19] WU S，WANG S. Information-theoretic outlier detection for large-scale categorical data[J]. IEEE Trans on Knowledge and Data Engineering，2013，25(3)：589-602.

**责任编辑**：郭红建