DOI:10.11918/202305067

改进 DPC 聚类算法的离群点检测与解释方法

周 玉,夏 浩,裴泽宣

(华北水利水电大学 电气工程学院, 郑州 450045)

摘 要: 为解决全局离群点检测方法无法对局部离群点进行检测,以及局部异常因子在面对大量局部离群点时性能下降的问题,利用 k 近邻(KNN)和核密度估计方法(KDE)提出一种基于改进快速搜索和发现密度峰值聚类算法(KDPC)的离群点检测与解释方法,该方法能够同时对数据点的全局和局部进行分析。首先,利用 k 近邻和核密度估计方法计算数据点的局部密度,代替传统 DPC 算法中根据截断距离计算的局部密度。其次,将数据点的 k 近邻距离之和作为全局异常值,并通过 KDPC 聚类算法计算簇密度以及数据点的局部异常值。最后,将数据点的全局与局部异常值进行乘积作为最终异常得分,选取异常得分最高的 Top-n 作为离群点,通过构建全局 - 局部异常值决策图对全局和局部离群点进行解释。利用人工数据集和 UCI 数据集进行实验并与 10 种常用离群点检测方法进行比较。结果表明,该方法对全局和局部离群点都有着较高的检测精度和检测性能,并且 AUC 方面受 k 值影响较小。同时,利用该方法对 NBA 球员数据进行分析讨论,进一步证明了该方法的实用性和有效性。

关键词: 离群点检测;聚类;密度峰值;k 近邻;核密度估计

中图分类号: TP181 **文献标志码:** A **文章编号:** 0367 - 6234(2024)08 - 0068 - 18

Improved outlier detection and interpretation method for DPC clustering algorithm

ZHOU Yu, XIA Hao, PEI Zexuan

(School of Electrical Engineering, North China University of Water Resources and Electric Power, Zhengzhou 450045, China)

Abstract: To address the limitatios of global outlier detection methods in detecting local outliers and the performance degradation of local anomaly factors in the presence of a large number of local outliers, this paper proposes an outlier detection and interpretation method based on an improved fast search and discovery density peak clustering algorithm (KDPC), utilizing k-nearest neighbor (KNN) and kernel density estimation (KDE) methods. This method enables simultaneous analysis of both global and local data points. Firstly, the local density of data points is calculated using the k-nearest neighbor and kernel density estimation methods instead of the local density based on the truncation distance in the traditional DPC algorithm. Secondly, the sum of the k-nearest neighbor distances of the data points is used as the global outlier and the cluster density as well as the local outliers of the data points are calculated by the KDPC clustering algorithm. Finally, the global and local outliers of the data points are multiplied as the final anomaly score. The Top-n data points with the highest anomaly score is selected as the outlier, and the global and local outliers are interpreted by constructing a global-local outlier decision diagram. Experiments were conducted using both artificial and UCI datasets and our method was compared with 10 commonly used outlier detection methods. The results show that our method achieves high detection accuracy and performance for both global and local outliers. Moreover, the AUC performance is minimally affected by the k-value. Additionally, our method is also used to analyze NBA player data, further demonstrating its practicality and effectiveness.

Keywords: outlier detection; clustering; density peaks; k-nearest neighbors; kernel density estimation

离群点被怀疑是由不同机制[1]产生的数据点,因不同于正常数据点而常被视作噪声点、也被认为是具有研究价值的点。目前离群点检测广泛应用于欺诈检测[2-3]、医疗处理[4-5]、环境卫生[6-7]、图像处理[8-9]、视频中异常事件检测[10-11]等。根据数据是否被标记,离群点检测方法分为有监督方法[12]、

收稿日期: 2023 - 05 - 23;**录用日期:** 2023 - 06 - 21;**网络首发日期:** 2024 - 06 - 05
网络首发地址: https://link.cnki.net/urlid/23.1235.T.20240604.1812.004
基金项目: 国家自然科学基金(U1504622, 31671580),河南省高等学校青年骨干教师培养计划项目(2018GGJS079)
作者简介: 周 玉(1979—),男,副教授,硕士生导师
通信作者: 周 玉,zhouyu_beijing@126.com

半监督方法[13]和无监督方法[14]。在难以获取数据标签的情况下,基于无监督的离群点检测成为常用的研究方法。无监督的离群点检测方法大致可分为:基于距离的离群点检测方法、基于密度的离群点检测方法和基于聚类的离群点检测方法[15-16]。

基于距离的方法能够快速地对离群点进行检测。Ramaswamy 等[17]提出 k 近邻(k-nearest neighbor, KNN)离群点检测方法,通过计算数据点与第 k 个最近邻居之间的距离,将距离大小作为数据点的异常值。对 KNN 方法的改进,Zhang 等[18]提出了一种基于局部距离的离群点检测方法(local distance-based outlier Factor, LDOF)。通过计算数据点的 k 近邻距离与 k 近邻数据点间的两两距离的比值来衡量数据点的离群程度。Yang 等[19]提出均值偏移的异常值检测方法(mean-shift outlier detector, MOD),通过计算数据点最近邻域的平均值对数据集进行均值偏移处理,然后计算数据点的偏移距离来确定离群程度。上述方法对全局离群点都有着不错的检测性能,但考虑到对局部异常值的检测效果差、难以适应不同密度的数据以及参数 k 对检测性能的显著影响。Xie 等[20]提出基于距离的局部引力离群点检测方法(local-gravitation outlier detection, LGOD),通过计算数据点的局部引力的变化率来得到数据点的异常程度。离群点的局部引力变化率高,而正常数据点的局部引力变化率低。

相比基于距离的离群点检测方法,基于密度的方法在局部异常值检测中表现良好,因为考虑邻域密度可以更好地反映数据的分布特征。Breuning 等[21]提出基于密度的局部异常值的离群点检测方法(local outlier Factor, LOF)。通过计算数据点的局部可达密度与其最近邻居密度的比率来计算离群因子,数据点的离群因子越大,表示离群程度越高。对于 LOF 算法的改进,Tang 等[22]提出了基于连通性的离群因子方法(connectivity-based outlier Factor, COF),将 LOF 的邻域计算方法改为增量计算,利用链距离提高了异常值检测性能。同样,Latecki 等[23]提出基于 LOF 的核密度估计方法(local density Factor, LDF)。对数据点的邻域进行局部核密度估计,然后通过计算数据点的局部密度与邻居的局部密度的比率来确定数据点的异常值。Tang 等[24]通过研究数据点的 k 最近邻、共享最近邻和反向近邻,并结合局部核密度估计方法对离群点进行检测,可以避免对不同密度数据的局部异常值的误判。张忠平等[25]提出了一种基于相对熵权密度离群因子的离群点检测算法。用熵权距离代替欧式距离,结合自然邻居对数据点局部进行高斯核密度估计,最后根据相对距离确定数据点的离群程度。此方法能够提高算法在低密度区域处理局部离群点的能力,但当局部离群点数量增多形成聚类时,离群点检测性能下降。

在基于聚类的离群点检测方法中,He 等[26]提出基于聚类的局部离群点检测方法(cluster-based local outlier Factor, CBLOF)。通过聚类方法将数据分成若干个聚类,给出倍数 β 来判别聚类的规模大小,将规模小的聚类视为离群类,并计算数据点的局部离群因子。利用聚类方法能对离群聚类进行筛除,提高离群点的检测精度。Al-Zoubl 等[27]首先使用 FCM 算法对数据集进行聚类并得到目标函数值。接着,通过剔除数据点后目标函数值变化量来判断数据点的离群程度。当数据形成不同密度的聚类时,该方法对局部离群点的检测性能变弱。对此,周玉等[28]首先利用 FCM 算法对数据集进行聚类。接着根据目标函数值剔除聚类中心附近的数据点。最后使用 LOF 算法计算剩余数据点的局部离群因子,提高局部离群点的检测精度。但上述方法受聚类算法限制,对任意形状数据集中的离群点检测性能不高。为此,张忠平等[29]提出了基于快速密度峰值聚类离群因子的离群点检测算法。首先使用 DPC 算法对数据集聚类,并根据数据点的 k 近邻平均距离定义了向心相对距离。接着,通过计算数据点的向心相对距离与 KNN 局部密度的比率得到离群因子。该方法通过聚类来分析每个簇的特征,根据离群因子的大小对离群点进行检测,同时能避免不同密度聚类对离群点检测带来的影响。

基于对全局离群点和局部离群点的考虑,本文提出基于 k 近邻和核密度估计方法的改进 DPC 聚类方法的离群点检测与解释方法。首先,通过计算数据点的 k 近邻距离,将 k 近邻距离之和作为数据点的全局异常值,并利用核密度估计方法计算数据点的局部密度。其次,根据局部密度得到数据点的相对距离,并对数据集进行聚类。再次,利用簇密度和局部密度的比率获取数据点的局部异常值。最后,将全局异常值与局部异常值相乘获取最终的异常得分,选取异常得分高的 Top-n 数据点作为离群点。通过构建全局 – 局部异常值决策图对离群点种类进行解释。

1　预备知识

1.1　DPC 聚类算法

快速搜索和发现密度峰值聚类算法(clustering

by fast search and find of density peaks, DPC) 是 Rodriguez 等[30]在 2014 年提出的一种聚类算法。构建关于密度与距离的决策图来确定聚类中心,即聚类中心密度最大且被密度小于它的数据点所包围,不同聚类中心间的距离较远。根据数据点归类于密度大且距离最近的簇的原则进行聚类。

DPC 算法中定义的两个参数如下。

定义 1 (局部密度):

$$\rho_i = \sum_i \chi(d_{ij} - d_c) \tag{1}$$

式中:当 $x<0$ 时,$\chi(x)=1$,否则 $\chi(x)=0$。d_{ij} 为数据点 i 与 j 之间的欧式距离,d_c 为截断距离。d_c 为超参数,通常被定义为:$d_c = p\%$,表示所有数据点的平均邻居数为数据总量的 $p\%$。避免数据点局部密度出现相同的情况,采用下式高斯函数来代替式(1),即

$$\rho_i = \sum_{i \neq j} \mathrm{e}^{-\left(\frac{d_{ij}}{d_c}\right)^2} \tag{2}$$

定义 2 (相对距离):

$$\delta_i = \min_{j:\rho_j > \rho_i} d_{ij} \tag{3}$$

计算数据点与密度大于它的最近数据点的距离。当数据点 i 的密度为数据集中最大时,式(3)不成立。此时,其相对距离为距离 i 最远点的距离,即 $\delta_i = \max_j d_{ij}$。利用式(2)、(3)获取数据点的局部密度和相对距离构建决策图来获取初始聚类中心。计算局部密度和相对距离的乘积来对聚类中心进行选取,即

$$\gamma = \rho \cdot \delta \tag{4}$$

图 1 是一组二维数据,图 1 中包含两个簇。构建局部密度 - 相对距离决策图如图 2 所示。局部密度大且相对距离大的数据点分布在决策图的右上角,这两个数据点是图 1 中两个簇中数据点分布最密集部分,将这两个数据点作为聚类中心。

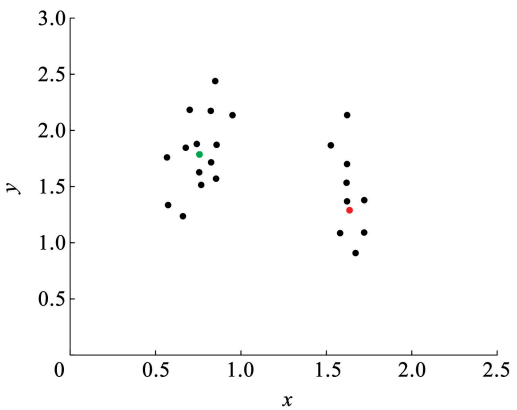


图 1 二维数据

Fig. 1 Two-dimensional data

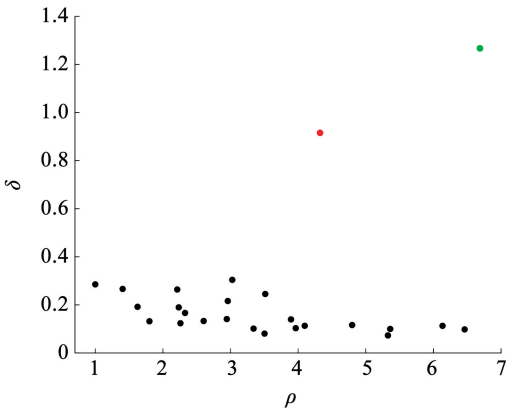


图 2 决策图

Fig. 2 Decision graph

确定聚类中心后,根据局部密度和相对距离的值将剩余数据点分配到密度更高且距离最近数据点所属类中。

1.2 k 近邻搜索方法

k 近邻搜索方法是通过计算数据集(N)中数据点间的欧氏距离,将距离值从小到大排列,找到数据点的 k 个最近邻居。计算公式如下:

$$d_k(x_i, x_j) = \sqrt{\sum_{m=1}^{d} (x_{im} - x_{jm})^2} \tag{5}$$

式中:$x_i, x_j \in N$,x_{im} 为第 i 个数据点的第 m 维,$x_i \neq x_j$,x_j 为 x_i 的第 k 近邻数据点,d 为维度个数。

如果两个数据点间的距离越小,那么这两个数据点越靠近。如果数据点的 k 近邻值很小,则该 $k+1$ 个数据点分布集中。为直观了解数据点的 k 近邻值与数据点分布之间的关系,随机生成二维数据如图 3 所示。图 3 中 P_1、P_2 为其中两个数据点,计算得到 P_1 与 P_2 的 k 近邻数据点,以 $k=5$ 为例。

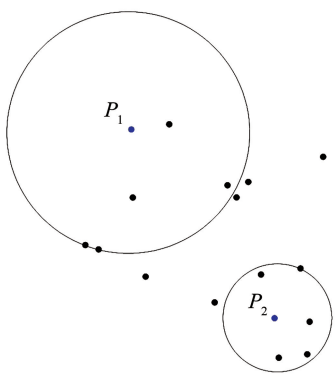


图 3 KNN 邻近分布示意($k=5$)

Fig. 3 Schematic representation of KNN neighbourhood distribution ($k=5$)

从图 3 中可知,P_1 周围的数据点分布相比 P_2 较为分散,使得 P_1 的 k 近邻数据点与 P_1 间的距离较远,即 $d_{k=5}(P_1) > d_{k=5}(P_2)$。通过 k 近邻距离值的大小可以判断数据点的分布情况,也可以对数据点的局部密度进行分析。

2　改进 DPC 聚类方法

当数据点形成不同密度的聚类时,根据传统 DPC 聚类算法难以确定簇密度小的聚类中心。基于此,利用 k 近邻和核密度估计方法计算数据点的局部密度,使得不同密度的簇都能有稳定的聚类中心。

2.1　局部密度

DPC 聚类方法中截断距离 d_c 的计算方法是:人为设定参数,使数据点的邻居平均数量约为数据集中总数据量的 1% ~2%。在簇密度大小有明显差异的数据集中,簇密度小的聚类中心很难被选取。图 4 为数据点形成密度差异大的 3 个聚类。

使用 DPC 方法构建决策图并对图 4 的数据集进行聚类。为解释数据点的聚类情况,分别对含有邻居平均数量约为数据集中总数据量的 2%、5%、10%、20%、50%、100%,以及对应的 d_c 下的决策图,如图 5 所示。

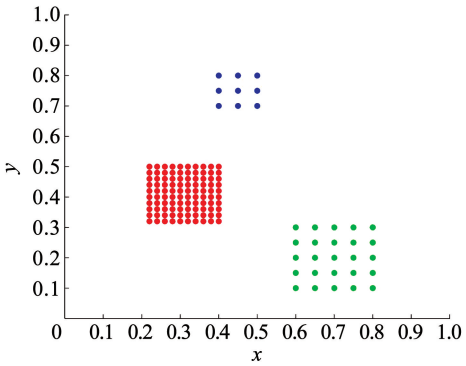
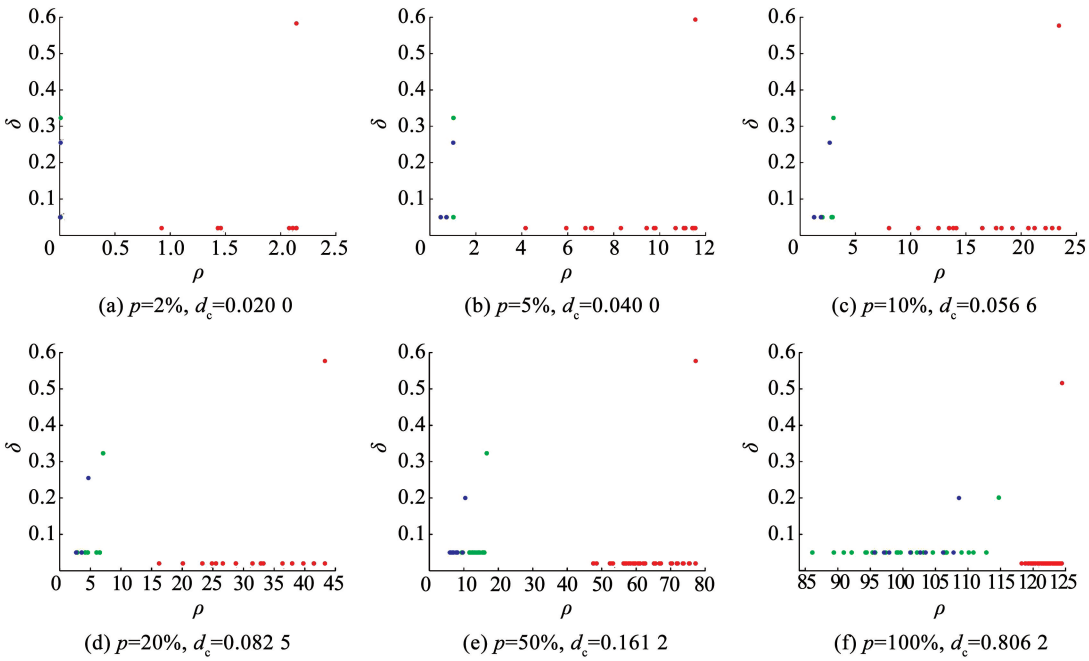


图 4　二维数据

Fig. 4　Two-dimensional data



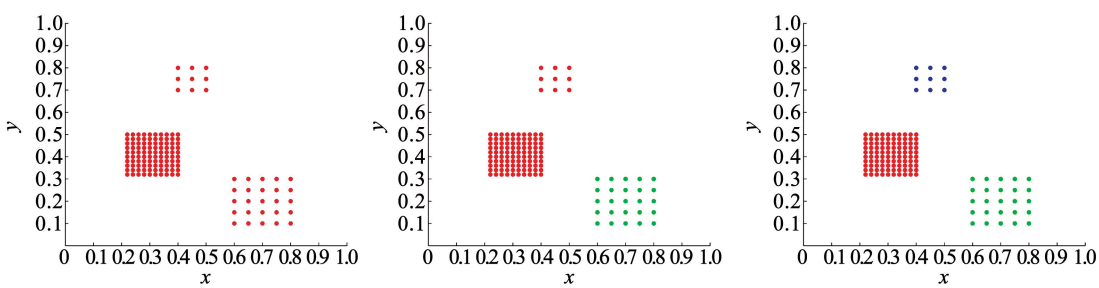
(a) p=2%, d_c=0.020 0　(b) p=5%, d_c=0.040 0　(c) p=10%, d_c=0.056 6

(d) p=20%, d_c=0.082 5　(e) p=50%, d_c=0.161 2　(f) p=100%, d_c=0.806 2

图 5　不同 d_c 下的决策图

Fig. 5　Decision diagram for different d_c values

从图 5 的结果看出,当数据点形成密度差异大的聚类时,较小的截断距离对应的决策图,无法识别密度小的聚类中心,进而形成错误的聚类。若要实现正确的聚类,需要正确选取 3 个聚类中心,即每个数据点含有邻居平均数量约为数据集中总数据量的 50% 以上。图 6 为选取不同数量的聚类中心所对应的聚类结果,c 为聚类中心个数。



(a) c=1　(b) c=2　(c) c=3

图 6　3 种不同的聚类结果

Fig. 6　Three different clustering results

2.2 核密度估计

为解决上述问题,计算数据点的 k 近邻距离并对数据点的局部密度进行核密度估计,让每个数据点都能有较为稳定的局部密度,进而能更加准确的聚类。利用高斯核函数对数据点的局部密度进行估计如式(6)所示。根据式(5)得到数据点的 k 近邻距离,并将其带入到式(6)并取平均值,得到式(7)。

$$K_{\text{Gaussian}}(x) = \frac{1}{(2\pi)^d}\exp\left(-\frac{\|x\|^2}{2}\right) \tag{6}$$

$$\rho(x_i) = \frac{1}{k}\sum_{x_j\in \text{KNN}(x_i)}\frac{1}{(2\pi)^d}\exp\left[-\frac{d_k(x_i,x_j)^2}{2}\right] \tag{7}$$

式中:$\|x\|$为 x 的范数,$d_k(x_i,x_j)$为数据点 x_i 与第 k 近邻的数据点 x_j 间的距离,d 为数据维度。

从式(6)、(7)可以看出 e^{-x} 为单调递减函数,当数据点的 k 近邻距离越大,局部密度越小,数据点分布越稀疏。为分析不同 k 下的决策图,如图 7 所示。

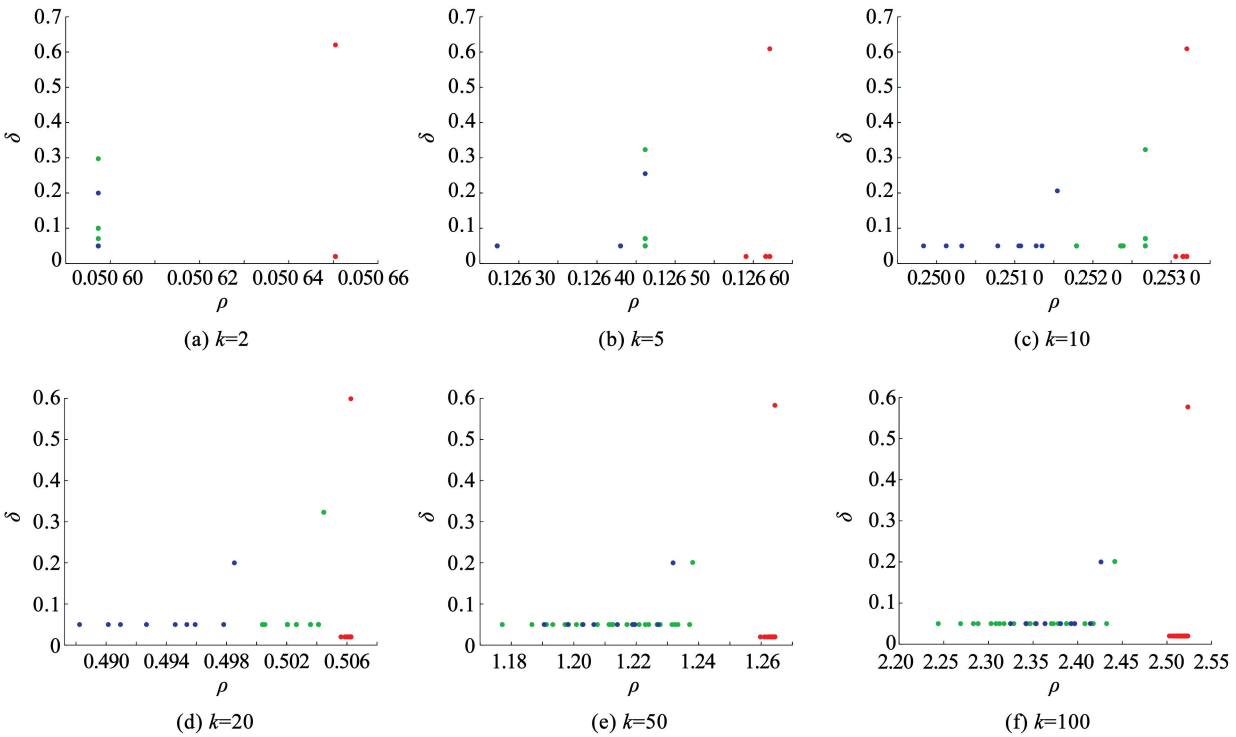


图 7 不同 k 下的决策图

Fig. 7 Decision graph of different values of k

从图 7 可以看出,当 $k=5$ 时,可以明显识别出 3 个不同簇中心在决策图的位置。利用 k 近邻的核密度估计方法能更加适应不同簇密度下的聚类。

3 基于改进 DPC 聚类算法的离群点检测方法

首先,利用 k 近邻方法对数据点的全局进行分析,计算数据点的全局异常值。其次,利用高斯核函数计算数据点的局部密度,并计算相对距离。再次,利用局部密度和相对距离值进行聚类,并计算聚类之后的每个簇的平均密度以及数据点的局部异常值。最后,结合全局异常值和局部异常值对离群点进行检测,并构建全局 – 局部异常值决策图对离群点进行解释。

3.1 全局异常值

当数据点分布密集时,数据点间的距离小,k 近邻的值小。相反,当数据点分布稀疏时,k 近邻的值大。离群点周围的数据分布比正常数据点周围的数据分布稀疏,使得离群点的 k 近邻数据点离群点较远。通过数据点的 k 近邻的值的大小能快速准确地检测出全局离群点。数据点的 k 近邻值的大小如式(8)所示。

$$S_{\text{Ges}}(x_i) = \sum_{x_j\in N_k(x_i)} d_k(x_i,x_j) \tag{8}$$

式中:$S_{\text{Ges}}(x_i)$ 为数据点 x_i 的全局异常值,$d_k(x_i,x_j)$ 为数据点 x_i 的第 k 近邻距离。S_{Ges} 越大,说明数据点间的密度越稀疏,则该数据点越有可能是离群点。当数据点形成密度差异大的聚类时,该方法只能检测簇密度小的周围的离群点,对整体数据集中的局部离群点检测性能下降。如图 8 所示,C_1、C_2 为两个密度相差较大的两个簇,O_1、O_2 为两个局部离群点。使用 KNN 方法无法对 C_1 簇中的 O_1 离群点进行检测。

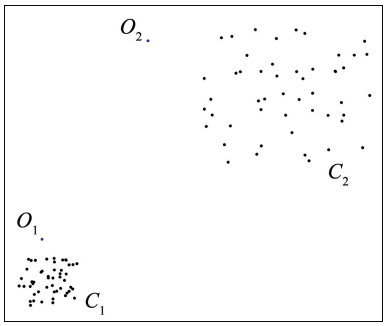


图 8　二维数据

Fig. 8　Two-dimensional data

3.2　局部异常值

考虑数据的局部离群信息,进而更加准确的对局部离群点进行检测。通过对数据集进行聚类获取若干个簇,计算簇的平均密度,从而计算数据点的局部异常值。通过聚类方法能快速获取每个数据点的局部异常值,降低计算复杂程度。

首先,根据式(5)得到数据点的 k 近邻值,利用式(6)、(7)得到数据点的局部密度。

其次,根据式(3)计算数据点的相对距离,使用快速搜索和发现密度峰值方法对数据进行聚类。根据式(9)计算簇的平均密度。

$$\bar{\rho}(l) = \frac{1}{k \cdot |N(l)|}\sum_{i=1}^{N(l)}\sum_{x_j \in \mathrm{KNN}(x_i)}\frac{1}{(2\pi)^d}\cdot \exp\left(-\frac{d_k(x_i,x_j)^2}{2}\right) \tag{9}$$

式中:$N(l)$为第 l 簇的数据量,x_i 为属于第 l 簇的第 i 个数据点。

关于数据点 x_i 的局部异常值解释:x_i 所在簇的平均密度与数据点 x_i 的核密度的比值,即

$$S_{\mathrm{Les}}(x_i) = \frac{\bar{\rho}(l)}{\rho(x_i)} \tag{10}$$

式中:$S_{\mathrm{Les}}(x_i)$ 为数据点 x_i 的局部异常值,$\bar{\rho}(l)$ 为第 l 簇的平均密度。

3.3　数据点异常得分

结合数据点的全局和局部的特点,将数据点的全局异常值与局部异常值进行相乘得到最终的异常得分,即

$$S_{\mathrm{KDPC}}(x_i) = S_{\mathrm{Les}}(x_i)\cdot S_{\mathrm{Ges}}(x_i) \tag{11}$$

式中:$S_{\mathrm{KDPC}}(x_i)$ 为数据点 x_i 的异常得分,选取数据点的异常得分最高的 Top-n 数据点视为离群点。

3.4　理论分析

本文将分别对全局异常值和局部异常值进行理论分析,证明离群点检测方法的有效性。

3.4.1　全局异常值

离群点是不同于正常特征属性的异常数据。直观来看,离群点远离正常数据点,数量少且分布稀疏。根据定义可以得到,离群点 x_{outlier} 的第 k 近邻值大于正常数据点 x_{normal} 的第 k 近邻值,则 $d_k(x_{\mathrm{outlier}}) > d_k(x_{\mathrm{normal}})$。得到:$\frac{S_{\mathrm{Ges}}(x_{\mathrm{outlier}})}{S_{\mathrm{Ges}}(x_{\mathrm{normal}})} > 1$。

证明　因为

$$d_k(x_{\mathrm{outlier}}) > d_k(x_{\mathrm{normal}}) \tag{12}$$

所以

$$\sum_{x_{\mathrm{outlier}}\in \mathrm{KNN}} d_k(x_{\mathrm{outlier}}) > \sum_{x_{\mathrm{normal}}\in \mathrm{KNN}} d_k(x_{\mathrm{normal}}) \tag{13}$$

即

$$\frac{S_{\mathrm{Ges}}(x_{\mathrm{outlier}})}{S_{\mathrm{Ges}}(x_{\mathrm{normal}})} = \frac{\sum_{x_{\mathrm{outlier}}\in \mathrm{KNN}} d_k(x_{\mathrm{outlier}})}{\sum_{x_{\mathrm{normal}}\in \mathrm{KNN}} d_k(x_{\mathrm{normal}})} > 1 \tag{14}$$

从 S_{Ges} 的大小可以得出结论:离群点的全局异常值大于正常数据点的全局异常值,可以通过全局异常值的大小对离群点进行分析。

3.4.2　局部异常值

考虑数据点的局部异常值,能更加精确地对离群点进行检测。因为 $d_k(x_{\mathrm{outlier}}) > d_k(x_{\mathrm{normal}})$,得到:$\frac{S_{\mathrm{Les}}(x_{\mathrm{outlier}})}{S_{\mathrm{Les}}(x_{\mathrm{normal}})} > 1$。

证明　令

$$d_k(x_{\mathrm{outlier}}) = \alpha_k d_k(x_{\mathrm{normal}}), \alpha_k > 1 \tag{15}$$

将式(15)带入式(6)可得

$$\frac{K_{\mathrm{Gaussian}}(x_{\mathrm{normal}})_k}{K_{\mathrm{Gaussian}}(x_{\mathrm{outlier}})_k} = \frac{\frac{1}{(2\pi)^d}\exp\left(-\frac{d_k(x_{\mathrm{normal}})^2}{2}\right)}{\frac{1}{(2\pi)^d}\exp\left(-\frac{\alpha_k^2 d_k(x_{\mathrm{normal}})^2}{2}\right)} = \exp\left(\frac{d_k(x_{\mathrm{normal}})^2}{2}(\alpha_k^2-1)\right) \tag{16}$$

因为 $\alpha_k > 1$,且指数函数 e^x 单调递增,那么:

$$\frac{K_{\mathrm{Gaussian}}(x_{\mathrm{normal}})_k}{K_{\mathrm{Gaussian}}(x_{\mathrm{outlier}})_k} = \exp\left(\frac{d_k(x_{\mathrm{normal}})^2}{2}(\alpha_k^2-1)\right) > e^0 = 1 \tag{17}$$

从式(17)可知,正常数据点的第 k 近邻的核密度估计值是大于离群点的第 k 近邻的核密度估计值,即

$$\frac{\rho(x_{\mathrm{normal}})}{\rho(x_{\mathrm{outlier}})} = \frac{\frac{1}{k}\sum_{x_{\mathrm{normal}}\in \mathrm{KNN}} K_{\mathrm{Gaussian}}(x_{\mathrm{normal}})_k}{\frac{1}{k}\sum_{x_{\mathrm{outlier}}\in \mathrm{KNN}} K_{\mathrm{Gaussian}}(x_{\mathrm{outlier}})_k} > 1 \tag{18}$$

从式(18)可以得知,正常数据点的局部密度大于离群点的局部密度。又因为离群点的数量远少于正常数据点的数量。在聚类完成后,第 l 簇的密度近似等于正常数据点的平均密度,即

$$\bar{\rho}(l) = \frac{1}{k \cdot |N(l)|}\sum_{i=1}^{N(l)}\sum_{x_j \in \mathrm{KNN}(x_i)}\frac{1}{(2\pi)^d}\exp\left(-\frac{d_k(x_i,x_j)^2}{2}\right) \approx \frac{1}{k \cdot |N(x_{\mathrm{normal}})|}\sum_{i=1}^{N(x_{\mathrm{normal}})}\sum_{x_j \in \mathrm{KNN}(x_i)}\frac{1}{(2\pi)^d}\exp\left(-\frac{d_k(x_i,x_j)^2}{2}\right) \approx \frac{1}{|N(x_{\mathrm{normal}})|}\sum_{i=1}^{N(x_{\mathrm{normal}})}\rho(x_{\mathrm{normal}}) \tag{19}$$

结合式(18)、(19)可得:

$$S_{\mathrm{Les}}(x_{\mathrm{normal}}) = \frac{\frac{1}{|N(x_{\mathrm{normal}})|}\sum_{i=1}^{N(x_{\mathrm{normal}})}\rho(x_{\mathrm{normal}})}{\rho(x_{\mathrm{normal}})} \tag{20}$$

$$S_{\mathrm{Les}}(x_{\mathrm{outlier}}) = \frac{\frac{1}{|N(x_{\mathrm{normal}})|}\sum_{i=1}^{N(x_{\mathrm{normal}})}\rho(x_{\mathrm{normal}})}{\rho(x_{\mathrm{outlier}})} \tag{21}$$

结合式(18)、(20)、(21)可得

$$\frac{S_{\mathrm{Les}}(x_{\mathrm{outlier}})}{S_{\mathrm{Les}}(x_{\mathrm{normal}})} = \frac{\rho(x_{\mathrm{normal}})}{\rho(x_{\mathrm{outlier}})} > 1 \tag{22}$$

根据式(22)可知,离群点的局部异常值大于正常数据点的局部异常值。根据局部异常值的大小可以对局部离群点进行分析。

3.4.3 异常得分

本文算法中,将全局异常值与局部异常值进行乘积作为最终的异常得分。方法的意义在于能够同时考虑全局与局部信息,且比其中的任意一种单独方法要好。

证明 根据式(14)、(22)可得:

$$\frac{S_{\mathrm{KDPC}}(x_{\mathrm{outlier}})}{S_{\mathrm{KDPC}}(x_{\mathrm{normal}})} = \frac{S_{\mathrm{Ges}}(x_{\mathrm{outlier}}) \cdot S_{\mathrm{Les}}(x_{\mathrm{outlier}})}{S_{\mathrm{Ges}}(x_{\mathrm{normal}}) \cdot S_{\mathrm{Les}}(x_{\mathrm{normal}})} > \frac{S_{\mathrm{Ges}}(x_{\mathrm{outlier}})}{S_{\mathrm{Ges}}(x_{\mathrm{normal}})} \tag{23}$$

$$\frac{S_{\mathrm{KDPC}}(x_{\mathrm{outlier}})}{S_{\mathrm{KDPC}}(x_{\mathrm{normal}})} = \frac{S_{\mathrm{Ges}}(x_{\mathrm{outlier}}) \cdot S_{\mathrm{Les}}(x_{\mathrm{outlier}})}{S_{\mathrm{Ges}}(x_{\mathrm{normal}}) \cdot S_{\mathrm{Les}}(x_{\mathrm{normal}})} > \frac{S_{\mathrm{Les}}(x_{\mathrm{outlier}})}{S_{\mathrm{Les}}(x_{\mathrm{normal}})} \tag{24}$$

下面将简单地给出一个证明,直观看出本文方法与仅考虑全局异常值方法之间的差异。

假设数据集中含有两个簇,两个簇中的数据点分布均匀,簇 1 的数据点密度高于簇 2 的数据点密度,即 $d_k(x_{\mathrm{normal}_2}) = \beta d_k(x_{\mathrm{normal}_1})$,其中 x_{normal_1} 与 x_{normal_2} 为簇 1 和簇 2 的正常数据点,且 $\beta > 1$。现在簇 1 中有一个局部离群点,局部离群点的 k 近邻距离是正常数据点的 k 近邻距离的 α 倍,且 $\alpha > 1$,即 $d_k(x_{\mathrm{outlier}}) = \alpha d_k(x_{\mathrm{normal}_1})$。首先计算数据点的全局异常值:

$$S_{\mathrm{Ges}}(x_{\mathrm{normal}_2}) = \sum_{x_{\mathrm{normal}_2} \in \mathrm{KNN}} d_k(x_{\mathrm{normal}_2}) = \beta \sum_{x_{\mathrm{normal}_1} \in \mathrm{KNN}} d_k(x_{\mathrm{normal}_1}) = \beta \cdot S_{\mathrm{Ges}}(x_{\mathrm{normal}_1}) \tag{25}$$

同理:

$$S_{\mathrm{Ges}}(x_{\mathrm{outlier}}) = \alpha \cdot S_{\mathrm{Ges}}(x_{\mathrm{normal}_1}) \tag{26}$$

结合式(25)、(26)可得

$$\frac{S_{\mathrm{Ges}}(x_{\mathrm{normal}_2})}{S_{\mathrm{Ges}}(x_{\mathrm{outlier}})} = \frac{\beta}{\alpha} \tag{27}$$

结果 1 根据式(27)可知,若要检测出数据点分布密度更大的簇 1 中的局部离群点,则 $\frac{\beta}{\alpha} < 1$,得 $\alpha > \beta$。

接着计算数据点的局部异常值,由式(22)可知:$\frac{S_{\mathrm{Les}}(x_{\mathrm{outlier}})}{S_{\mathrm{Les}}(x_{\mathrm{normal}_1})} > 1$,令 $S_{\mathrm{Les}}(x_{\mathrm{outlier}}) = \lambda \cdot S_{\mathrm{Les}}(x_{\mathrm{normal}_1})$ 其中 $\lambda > 1$。又因为根据式(19)、(20)可知,在数据点分布密度均匀的若干个簇中,正常数据点的局部异常值近似相等,即

$$S_{\mathrm{Les}}(x_{\mathrm{outlier}}) = \lambda \cdot S_{\mathrm{Les}}(x_{\mathrm{normal}_1}) \approx \lambda \cdot S_{\mathrm{Les}}(x_{\mathrm{normal}_2}) \tag{28}$$

结合式(27)、(28)可得

$$\frac{S_{\mathrm{KDPC}}(x_{\mathrm{normal}_2})}{S_{\mathrm{KDPC}}(x_{\mathrm{outlier}})} = \frac{S_{\mathrm{Ges}}(x_{\mathrm{normal}_2}) \cdot S_{\mathrm{Les}}(x_{\mathrm{normal}_2})}{S_{\mathrm{Ges}}(x_{\mathrm{outlier}}) \cdot S_{\mathrm{Les}}(x_{\mathrm{outlier}})} = \frac{\beta}{\lambda\alpha} \tag{29}$$

结果 2 此时需要检测出数据点分布密度更大的簇 1 中的局部离群点,则 $\frac{\beta}{\lambda\alpha} < 1$,得 $\alpha > \frac{\beta}{\lambda}$。

统计这两次的结果见表 1。从表 1 中可知,$\alpha_1 > \alpha_2$。说明在没有乘以局部异常值之前,需要更大的 α 才能检测出局部离群点。显然离群点的分布是确定的,只有降低检测出局部离群点所需要的 α 才是提高离群点检测精度的关键。综上所述,KDPC 方法能在检测全局离群点的同时,兼顾对局部离群点的考虑。

表 1 结果对比

Tab. 1 Comparison of results

参数 α	全局异常值(S_{Ges})	异常得分(S_{KDPC})
α 的取值范围	$\alpha_1 > \beta$	$\alpha_2 > \frac{\beta}{\lambda}$

3.5 计算复杂度分析

KDPC 计算复杂度的分析如下:

1)使用 KD 树搜索 k 个最近邻居的时间是 $O(N\log N)$,其中 N 为数据集中的样本数量;

2)计算 $S_{Ges}(x_i)$ 和 $S_{Les}(x_i)$ 的复杂度是 O(N)。

综上所述,计算 KDPC 的复杂度为 O(Nlog N) + 2O(N) = O(Nlog N),和 KNN、LOF 方法拥有着相同的计算复杂度。但本文利用 DPC 聚类方法,可以得到聚类后的数据点种类以及分布情况。在能够检测全局和局部离群点的情况下,对数据点进行分析解释。

3.6　算法步骤

输入　数据集、近邻参数 k。

输出　数据点的异常得分。

Step1　根据式(5)计算数据点的 k 近邻距离,并根据式(8)得到数据点的全局异常值。

Step2　根据式(6)、(7)对数据点进行核密度估计,计算数据点的局部密度 ρ。

Step3　根据式(3)计算数据点的相对距离 δ。

Step4　依据 ρ 和 δ 选择聚类中心。

Step5　将剩余数据分配到密度更高且距离最近数据点所属类中。

Step6　根据式(9)、(10)计算每个簇的平均密度及数据点的局部异常值。

Step7　结合全局异常值和局部异常值,根据式(11)得到数据点最终异常得分。

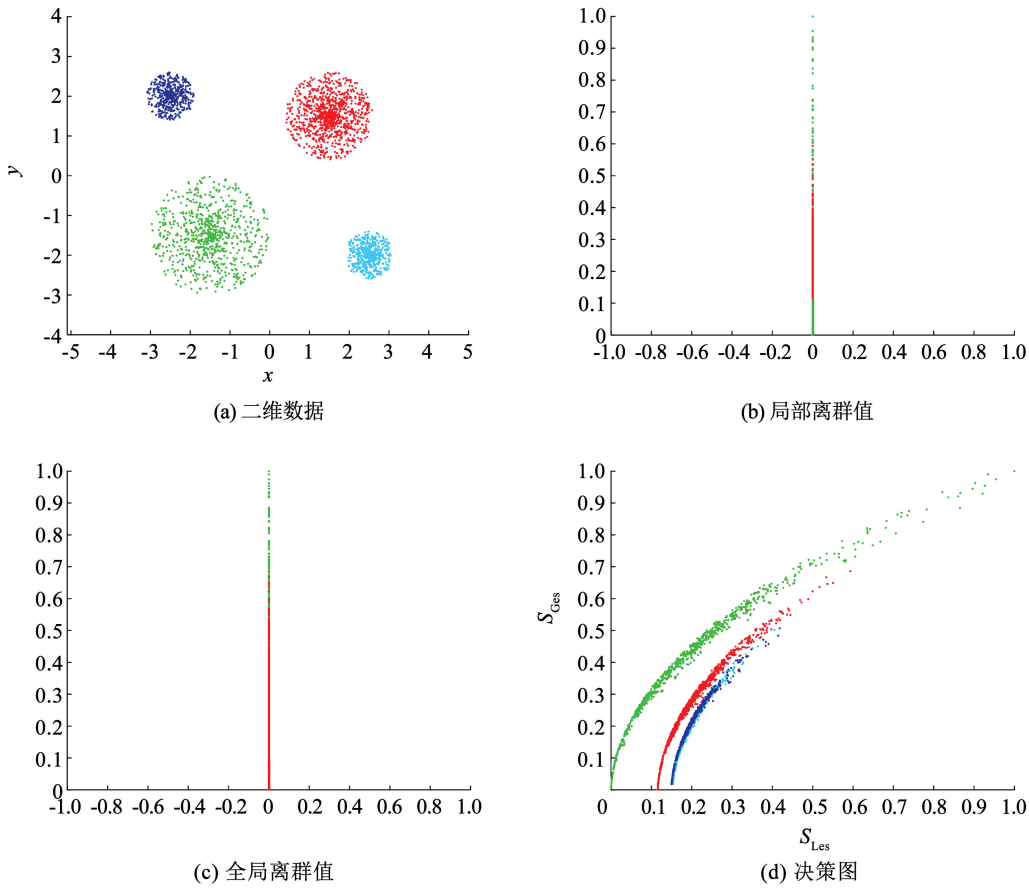
Step8　构建全局 – 局部异常值决策图,对数据点离群程度进行解释。

4　离群点种类的解释

本文将式(8)与式(10)得到的全局异常值与局部异常值进行归一化,以纵坐标为全局异常值、以横坐标为局部异常值,构建全局 – 局部离群值决策图,对在不同数量和种类下的离群点进行解释。

4.1　数据集中无离群点

正常数据形成的 4 个聚类如图 9(a)所示,正常数据点的局部离群值、全局离群值大小接近,在决策图中数据点分布均匀连续如图 9(d)所示。可以通过决策图中的位置了解每个簇数据点的分布情况,例如,分布在决策图左上方的数据点有着更小的局部离群值和更大的全局离群值,这类簇数据点分布会更加稀疏。



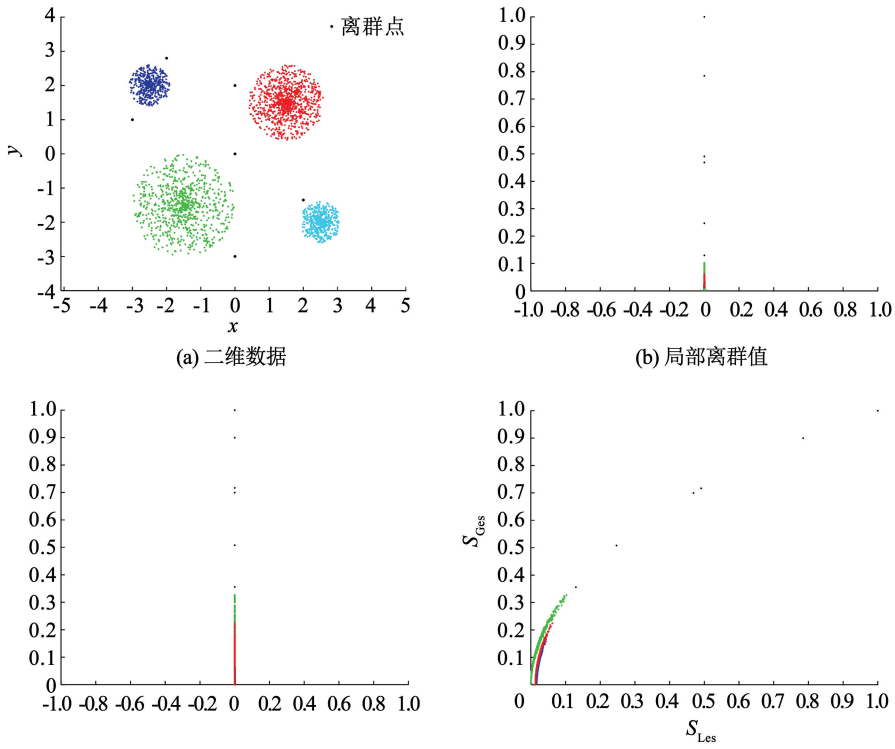
(a) 二维数据　(b) 局部离群值　(c) 全局离群值　(d) 决策图

图 9　数据集中无离群点

Fig. 9　Dataset contains no outliers

4.2　数据集中含少量局部离群点

在图 9(a)数据基础上增加部分局部离群点如图 10(a)所示。局部离群点相比正常数据点有较大的局部异常值与全局异常值。在图 10(d)中,离群点会更加远离密集数据点在空间中分布稀疏。



(a) 二维数据 (b) 局部离群值

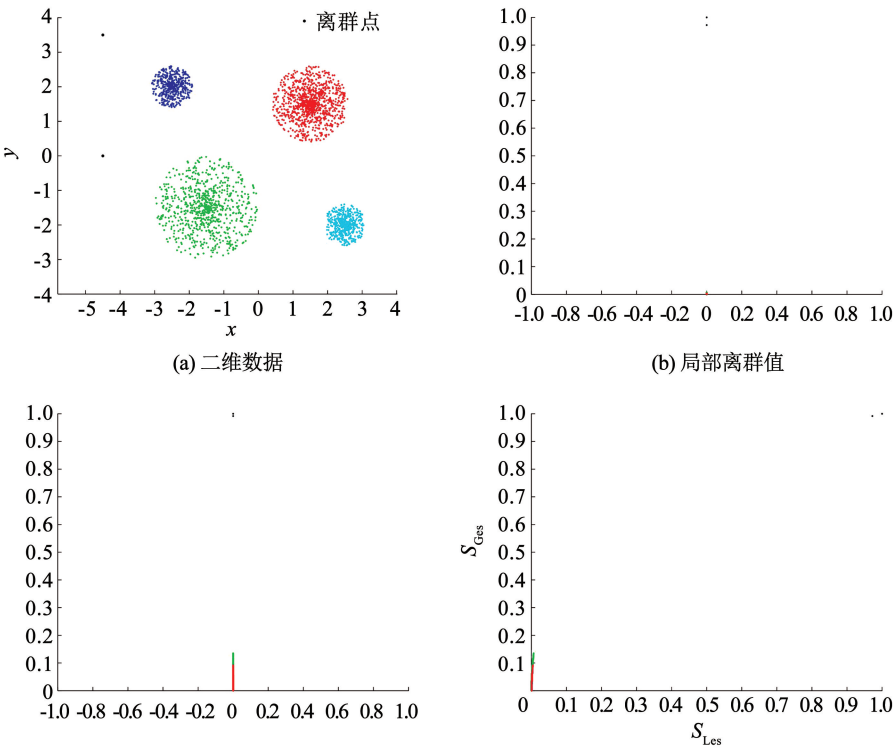
(c) 全局离群值 (d) 决策图

图 10 数据集中含少量局部离群点

Fig. 10 Dataset contains few local outliers

4.3 数据集中含少量全局离群点

增加少量全局离群点如图 11(a)所示。全局离群点比局部离群点更加远离正常数据。相比局部离群点有更大的全局异常值和局部异常值。与正常数据点相比有着悬殊的异常值差距,在归一化之后,正常数据点在决策图的最左上角,全局离群点位于最右上角如图 11(d)所示。



(a) 二维数据 (b) 局部离群值

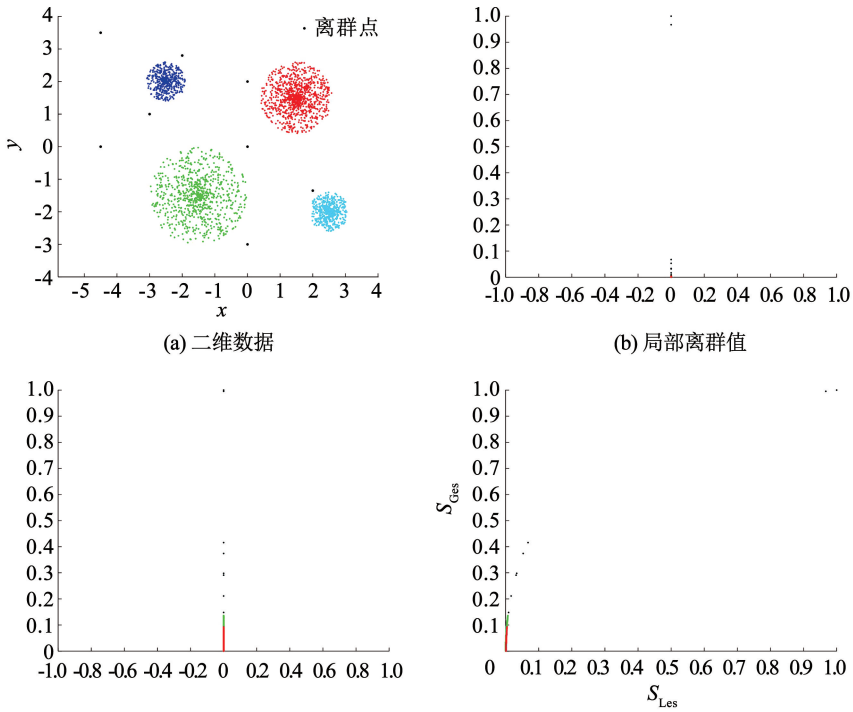
(c) 全局离群值 (d) 决策图

图 11 数据集中含少量全局离群点

Fig. 11 Dataset contains few global outliers

4.4　数据集中同时存在局部离群点和全局离群点

图 12(a)包含全局离群点和局部离群点。决策图中同时含有全局离群点和局部离群点时,由于归一化的原因局部离群点会更加靠近正常数据点,且分布较为稀疏。全局离群点位置不受影响,处于决策图坐标点(1,1)附近。



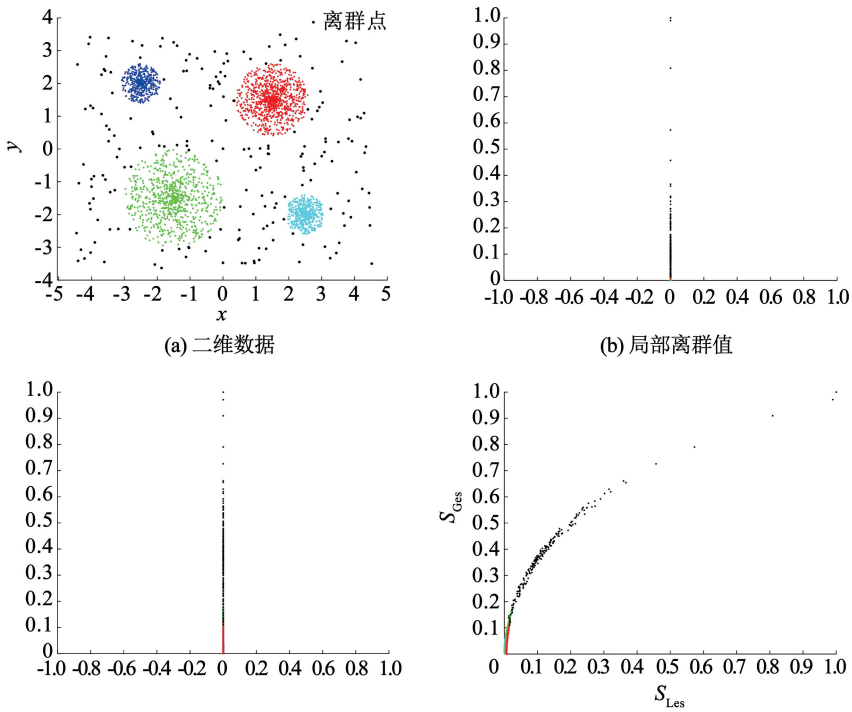
(a) 二维数据　(b) 局部离群值　(c) 全局离群值　(d) 决策图

图 12　数据集中存在局部离群点和全局离群点

Fig. 12　Local and global outliers exist in the dataset

4.5　数据集中局部离群点与全局离群点数量增多

在图 13(a)中,当局部离群点与全局离群点数量增加且不存在明显界限时,决策图如图 13(d)所示。靠近连续数据点的松散数据点视为局部离群点,靠近决策图坐标点(1,1)的数据点可视为全局离群点。



(a) 二维数据　(b) 局部离群值　(c) 全局离群值　(d) 决策图

图 13　数据集中存在较多的局部离群点和全局离群点

Fig. 13　A large number of local and global outliers exist in the dataset

5 实验分析

5.1 人工数据集

为解释本文方法的离群点检测性能,通过人工数据集对离群点进行检测,人工数据集参数见表 2。

表 2 人工数据集

Tab. 2 Artificial datasets

数据名称	数据量	簇数	维度	离群点数量
Data1	700	3	2	20
Data2	320	3	2	20
Data3	650	2	2	50
Data4	316	3	2	10
Data5	5 650	4	2	150
Aggregation	803	7	2	15

首先,将数据进行[0,1]归一化处理,计算数据点的 k 近邻距离,得到数据点的全局异常值。其次,利用高斯核函数对数据点的局部核密度值进行计算,并计算数据点的高密度最小距离的值,通过决策图确定聚类中心。再次,利用快速搜索和发现密度峰值聚类方法对数据集进行聚类并根据式(10)计算数据点的局部异常值。最后,将全局异常值与局部异常值进行相乘获取数据点的异常得分,选取异常值最高的 Top-n 个数据点作为离群点,数据形状及离群点分布如图 14 所示。

利用全局异常值和局部异常值构建决策图,对数据点的种类进行解释,如图 15 所示。位于决策图的左下部分密集的数据点可视为正常数据点;靠中间部分的数据点较为分散,这些数据点可视为局部离群点;在右上角的数据点,全局异常值与局部异常值均为最大,该类型的数据点可视为全局离群点。

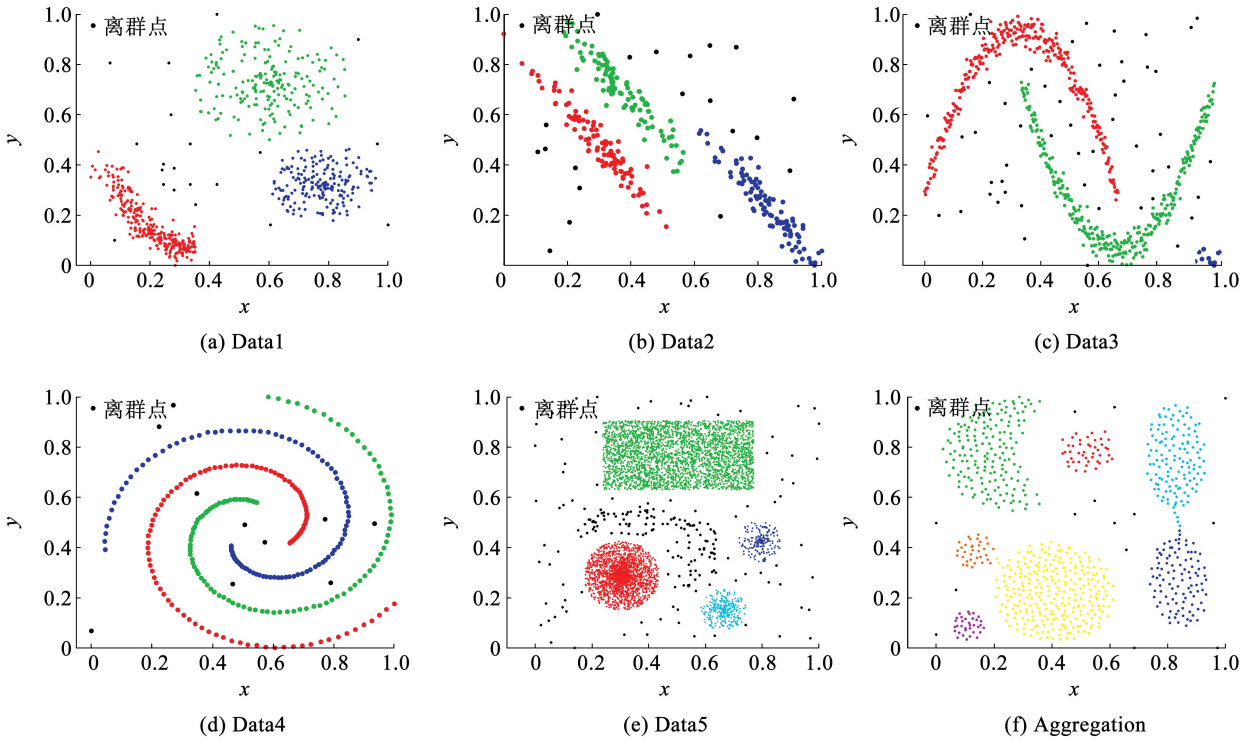


图 14 二维人工数据集

Fig. 14 Two-dimensional artificial datasets

关于离群点检测指标,用准确率(Precise)、召回率(Recall)、F_1 以及 AUC 值(ROC 曲线下降面积)[31-32]来评估该算法的离群点检测性能,即:

$$P = \frac{T_P}{T_P + F_P} \tag{30}$$

$$R = \frac{T_P}{T_P + F_N} \tag{31}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{32}$$

式中:T_P 为算法检测到真实的离群点数量,F_P 为算法把正常数据错分成离群点的数量,F_N 为把离群点识别成正常数据点的数量,T_N 为把正常数据点识别为正常数据点的数量。Precise、Recall、AUC 以及 F_1 的取值范围为[0,1],数值越大离群点检测性能越好。

关于阈值的选择,在人工数据集中选取最终异常值得分最高的 Top-n 个数据点作为离群点。检测指标见表 3。为了将本文算法与其他离群点检测方法进行对比,表 4 中包括 10 种常见的离群点检测方法。接着,通过计算各个方法的 AUC 值来对比离群点检测性能,见表 5。

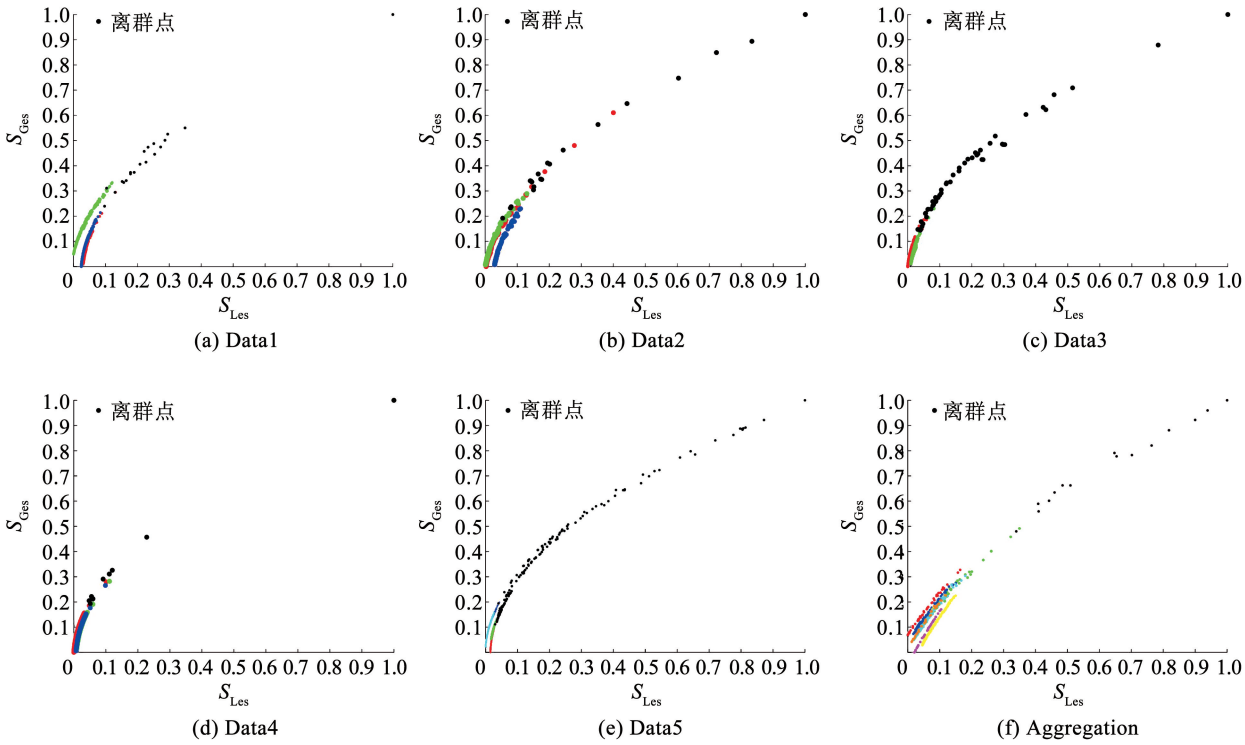


图 15　决策图

Fig. 15　Decision graph

表 3　离群点检测的性能指标

Tab. 3　Performance indicators of outlier detection

数据集	Top-n	Precise	recall	F_1	AUC	数据集	Top-n	Precise	recall	F_1	AUC
Data1	20	0.900 0	0.900 0	0.900 0	0.999 5	Data4	10	1.000 0	1.000 0	1.000 0	1.000 0
Data2	20	0.900 0	0.900 0	0.900 0	0.996 5	Data5	150	0.913 3	0.913 3	0.913 3	0.999 8
Data3	50	0.960 0	0.960 0	0.960 0	1.000 0	Aggregation	15	0.933 3	0.933 3	0.933 3	1.000 0

表 4　比较方法

Tab. 4　Comparison methods

方法	基本原理	出版者,出版年	方法	基本原理	出版者,出版年
LOF[21]	密度	ACM SIGMOD, 2000	ABOD[33]	角度	KDD, 2008
KNN[17]	距离	ACM SIGMOD, 2000	LDOF[18]	距离	AKDDM, 2009
COF[22]	密度	AKDDM, 2002	FCM[27]	聚类	WSEAS, 2010
CBLOF[26]	密度	Pattern Recognition Letters, 2003	IFOREST[34]	二叉树	TKDD,2008
LDF[23]	密度	MLDM, 2007	MOD[19]	均值漂移	Pattern Recognition, 2021

表 5　不同离群点检测方法的 AUC 值

Tab. 5　AUC values of different outlier detection methods

数据集	KDPC	LOF	KNN	COF	CBLOF	LDF	ABOD	LDOF	FCM	IFOREST	MOD
Data1	**0.999 5**	0.993 1	0.999 4	0.996 2	0.922 5	0.998 4	0.984 6	0.946 2	0.920 5	0.925 8	0.998 7
Data2	0.996 5	0.996 0	0.996 3	0.993 8	0.901 0	**0.996 8**	0.996 3	0.990 2	0.915 2	0.949 5	0.996 5
Data3	**1.000 0**	0.999 2	0.999 9	0.992 9	0.679 4	0.998 8	0.999 9	0.966 5	0.703 3	0.820 0	0.998 1
Data4	**1.000 0**	0.905 5	0.995 4	1.000 0	0.560 9	1.000 0	1.000 0	0.996 4	0.551 8	0.610 6	0.999 0
Data5	**0.999 8**	0.999 7	0.999 5	0.977 8	0.954 9	0.949 4	0.946 6	0.953 2	0.960 2	0.936 5	0.999 2
Aggregation	**1.000 0**	0.999 8	1.000 0	1.000 0	0.985 8	1.000 0	1.000 0	0.975 3	0.717 3	0.957 9	1.000 0

通过表 5 得到的 AUC 值,得出以下结论:

1)通过 Data1 ~ Data4 可知,离群点检测方法均不错。但由于聚类算法对数据集形状的要求较为苛刻,所以任意形状的数据集(Data3、Data4)会导致 CBLOF 和 FCM 的离群点检测性能下降。

2)Data5 数据集中存在明显的不平衡情况,且存在离群聚类的情况。在此情况下,KDPC 方法优于其他离群点检测方法。

3)Aggregation 也存在不平衡的情况,因此 FCM 的离群点检测性能受到影响,其他离群点检测方法都有较好的检测性能。

5.2 UCI 数据集

本文将用 UCI 数据集进行离群点检测方法的对比实验,验证 KDPC 算法的性能,表 6 中含有 13 种数据集。

计算各个离群点检测方法的 AUC 值,结果见表 7。为直观了解不同 k 下各个离群点检测性能,绘制 k = 2 ~ 100 下各个离群点检测方法的 AUC 曲线图,如图 16 所示。KDPC 算法在不同 k 的离群点检测性能具有稳定性,且在较低的 k 下,就有较高的离群点检测精度。算法采用 MatlabR2020a 编写,实验环境为:AMDR7 3.2 GHz CPU, 8.00 GB 内存,Windows11 操作系统。

表 6 UCI 数据集

Tab. 6 UCI datasets

数据集	维度	类数	离群类	数据量	离群点数量
Iris	4	3	3	105	5
Wine	13	3	2	115	8
Seeds	7	3	2	148	8
WPBC	33	2	2	161	10
Ionosphere	34	2	1	351	126
WBC	9	2	2	464	20
WDBC	30	2	2	569	212
Vowel	3	6	1	809	10
Waveform	21	3	1	3 443	100
Robot navigation	24	4	4	5 148	20
Page Blocks	10	2	2	5 473	560
Annthyroid	6	2	2	7 200	534
Pen Digits	16	10	1	9 879	30

表 7 UCI 数据集的 AUC 对比

Tab. 7 Comparison of AUC for the UCI dataset

数据集	KDPC	LOF	KNN	COF	CBLOF	LDF	ABOD	LDOF	FCM	IFOREST	MOD
Iris	**1.000 0**	0.996 0	1.000 0	0.990 0	1.000 0	0.986 0	0.964 0	0.992 0	1.000 0	0.981 0	0.994 0
Wine	**1.000 0**	0.995 3	0.993 0	0.988 3	0.982 5	1.000 0	0.952 1	0.959 1	0.974 3	0.953 7	0.991 8
Seeds	0.999 1	**1.000 0**	0.999 1	0.997 3	0.995 5	0.996 4	0.990 2	0.993 8	0.996 0	0.996 8	0.999 1
WPBC	0.691 4	0.702 6	**0.707 3**	0.688 1	0.541 7	0.696 7	0.694 0	0.703 3	0.561 6	0.644 3	0.680 1
Ionosphere	**0.935 4**	0.921 8	0.933 1	0.898 7	0.750 2	0.919 4	0.806 9	0.893 5	0.541 4	0.848 7	0.931 3
WBC	**1.000 0**	1.000 0	1.000 0	0.956 2	0.990 5	0.729 2	0.979 5	0.911 8	0.989 0	0.997 3	0.960 4
WDBC	0.892 0	**0.960 3**	0.922 9	0.553 2	0.733 3	0.552 3	0.839 9	0.556 1	0.609 9	0.811 5	0.735 4
Vowel	0.927 0	**0.942 8**	0.926 2	0.801 9	0.614 1	0.859 9	0.905 6	0.630 8	0.479 4	0.767 1	0.894 1
Waveform	0.783 8	**0.809 3**	0.787 8	0.754 2	0.708 3	0.784 5	0.746 6	0.725 8	0.680 0	0.692 7	0.771 4
Robot navigation	0.886 4	**0.889 8**	0.881 8	0.780 2	0.618 4	0.692 4	0.864 1	0.684 5	0.504 1	0.825 8	0.865 5
Page Blocks	0.767 1	0.763 0	0.841 4	0.784 9	**0.892 0**	0.824 6	0.703 9	0.830 9	0.537 0	0.889 8	0.783 9
Annthyroid	0.763 0	0.743 9	0.756 0	0.733 4	0.643 3	0.687 4	0.758 7	0.757 4	0.504 4	**0.818 6**	0.756 2
Pen Digits	0.998 0	0.996 2	0.997 1	0.976 8	0.981 5	0.979 5	0.950 7	0.798 5	0.974 5	0.973 9	**0.998 1**

5.3 NBA 球员数据集

从 NBA 中国官方网站(https://china.nba.cn/statistics/)获取 2021 — 2022 年季后赛的 50 位球员的数据,得到球员的 14 个方面的信息:场均得分、场均篮板、场均助攻、出场时间、进攻效率、投篮命中率、三分命中率、罚球命中率、进攻、防守、场均抢断、场均盖帽、失误、犯规。利用这 14 个方面信息计算球员的全局异常值与局部异常值,并构建决策图,如图 17 所示。

通过决策图发现,有 4 个点远离其他大多数球员数据,对这 4 个数据点进行标号,分别是 1 号扬尼斯·安特托昆博、3 号尼古拉·约基奇、45 号约纳斯·瓦兰丘纳斯和37 号贾伦·杰克逊,相应的球员及数据见表 8。为方便对比,计算 50 个球员的平均值。

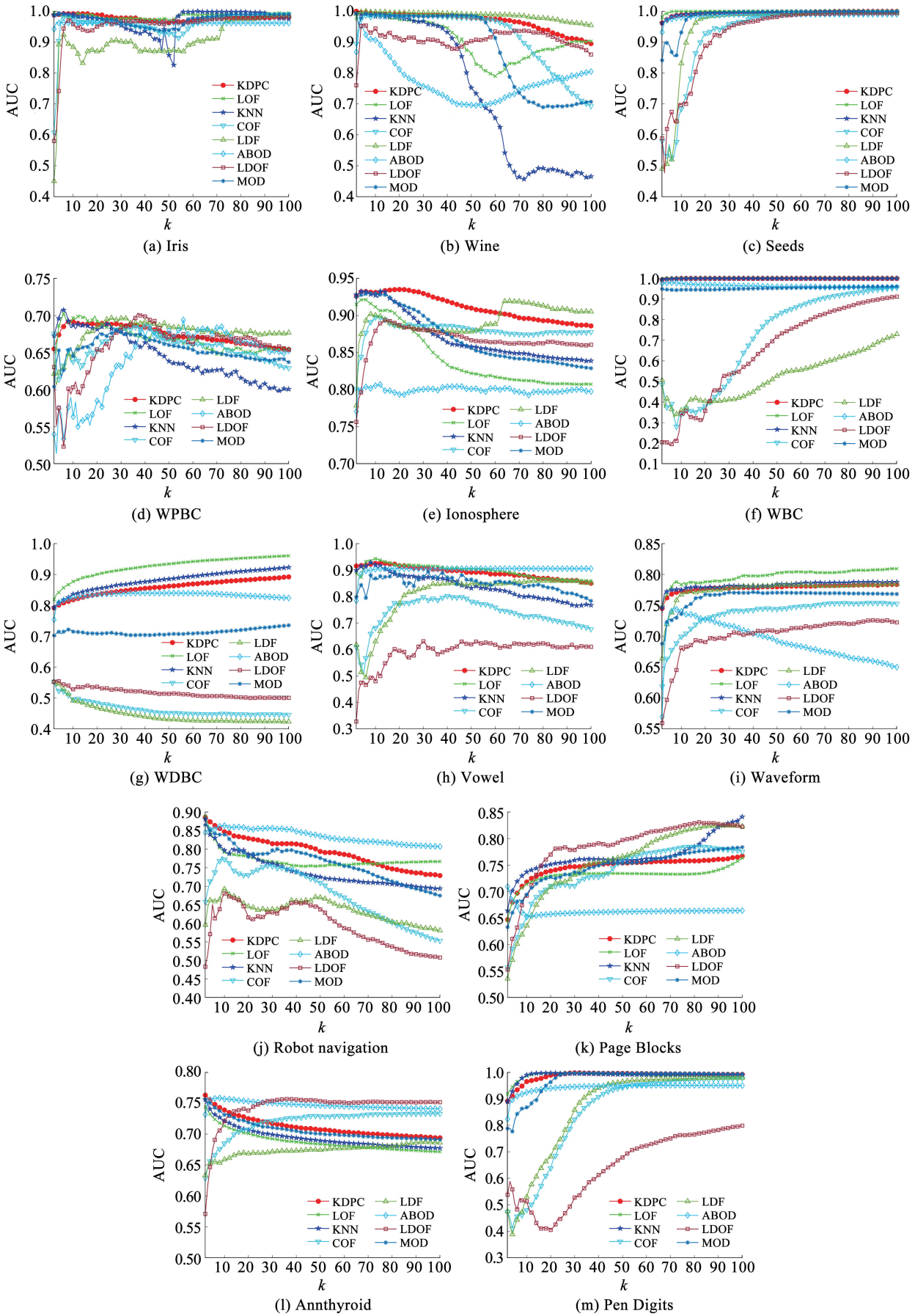


图 16　离群点检测方法在不同 k 下的 AUC 值

Fig. 16　AUC values of outlier detection methods at different k values

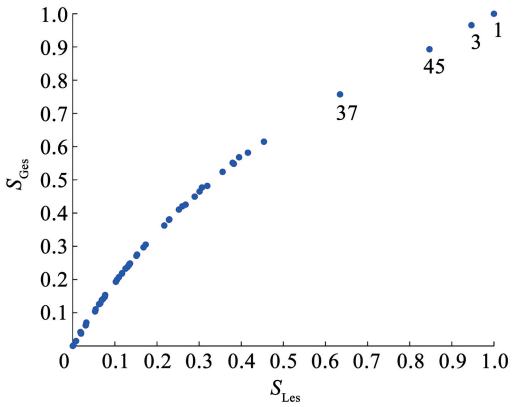


图 17 NBA 球员数据决策图

Fig. 17 NBA player data decision graph

从表 8 数据中可以分析：

1)1 号球员最为离群，可视为全局离群点。原因在于他的场均得分、场均篮板、防守、失误都远高于平均值，三分命中率和罚球命中率也低于平均水平。

2)3 号球员可视为全局离群点。他的场均得分、场均篮板、进攻效率、防守、失误都高于平均水平。

3)45 号球员也可视为全局离群点。他离群的主要原因在于有远高于平均水平的场均篮板、进攻和防守，三分命中率也远低于平均水平。

4)37 号球员可视为局部离群点，他能离群的主要原因在于有远低于平均水平的场均助攻，但他的场均盖帽是高于平均水平的。从上述分析可以发现，成为离群点的原因在于有若干方面是明显异于平均值。这些球员有他们的优点，也有他们的短板。然而，NBA 球员都是篮球界的精英，从全部属性分析球员，往往不能挖掘全部的信息。接下来本文将从单个属性对每个球员进行分析，找出该方面最具突出的球员，决策图如图 18 所示。通过对决策图的观察，分析每个特征属性，找到对应的离群球员见表 9。

表 8 部分 NBA 球员数据

Tab. 8 Selected NBA players statistics

序号	场均得分	场均篮板	场均助攻	出场时间	进攻效率	投篮命中率/%	三分命中率/%	罚球命中率/%	进攻	防守	场均抢断	场均盖帽	失误	犯规
1	31.7	14.2	6.8	37.3	34.3	49.1	22.0	67.9	2.2	12.0	0.7	1.3	4.5	3.6
3	31.0	13.2	5.8	34.2	37.8	57.5	27.8	84.0	3.4	9.8	1.6	1.0	4.8	4.0
45	14.5	14.3	3.0	29.1	23.8	48.5	16.7	76.9	5.5	8.8	0.7	0.2	2.0	2.8
37	15.4	6.8	0.9	27.7	15.8	37.8	37.5	75.5	2.2	4.6	0.8	2.5	1.4	4.4
均值	20.1	6.1	4.2	35.7	20.1	46.2	34.5	81.6	1.2	4.9	1.0	0.6	2.6	2.9

从表 9 数据中可以分析：

1)场均得分。1 号球员扬尼斯·安特托昆博和 3 号球员尼古拉·约基奇可视为全局离群点，因为他们的场均得分分别为 31.7 和 31.0。

2)场均篮板。45 号球员约纳斯·瓦兰丘纳斯和 1 号球员扬尼斯·安特托昆博可视为全局离群点，3 号球员尼古拉·约基奇和 23 号球员尼古拉·武切维奇可视为局部离群点。场均篮板分别为：14.3、14.2、13.2 和 12.4。

3)场均助攻。6 号球员贾·莫兰特可视为全局离群点，28 号球员詹姆斯·哈登和 32 球员克里斯·保罗可视为局部离群点。场均助攻分别为：9.8、8.6 和 8.3。

4)出场时间。46 号球员博格丹·博格达诺维奇出场时间为 26.7，可视为局部离群点。

5)进攻效率。3 号球员尼古拉·约基奇可视为全局离群点，1 号球员扬尼斯·安特托昆博可视为局部离群点，进攻效率分别为：37.8 和 34.3。

6)投篮命中率。30 号球员德安德烈·艾顿可视为全局离群点，39 号球员特雷·杨和 41 号球员巴姆·阿德巴约可视为局部离群点，投篮命中分别为：64.0%、31.9% 和 59.4%。

7)三分命中率。21 号球员德马尔·德罗赞视为全局离群点，三分命中率为 0%。

8)罚球命中率。在这一属性中，作为职业球员罚球命中都很出色，没有离群的球员。

9)进攻。45 号球员约纳斯·瓦兰丘纳斯可视为全局离群点，3 号球员尼古拉·约基奇可视为局部离群点，进攻分别为：5.5 和 3.4。

10)防守。1 号球员扬尼斯·安特托昆博视为全局离群点，防守值为 12.0。

11)场均抢断。2 号球员卢卡·东契奇、4 号球员吉米·巴特勒和 6 号球员贾·莫兰特可视为局部离群点。场均抢断分别为：1.8、2.1 和 2.0。

12)场均盖帽。37 号球员贾伦·杰克逊可视为全局离群点，17 号球员卡尔·安东尼唐斯可视为局部离群点，场均盖帽为：2.5 和 2.0。

13)失误。39 号球员特雷·杨可视为全局离群点，8 号球员凯文·杜兰特和 44 号球员克里斯·米德尔顿可视为局部离群点，失误分别为：6.2、5.3 和 5.5。

14)犯规。37 号球员贾伦·杰克逊可视为全局离群点，4 号球员吉米·巴特勒和 17 号球员卡尔·安东尼唐斯可视为局部离群点，犯规数：4.4、1.5 和 4.2。

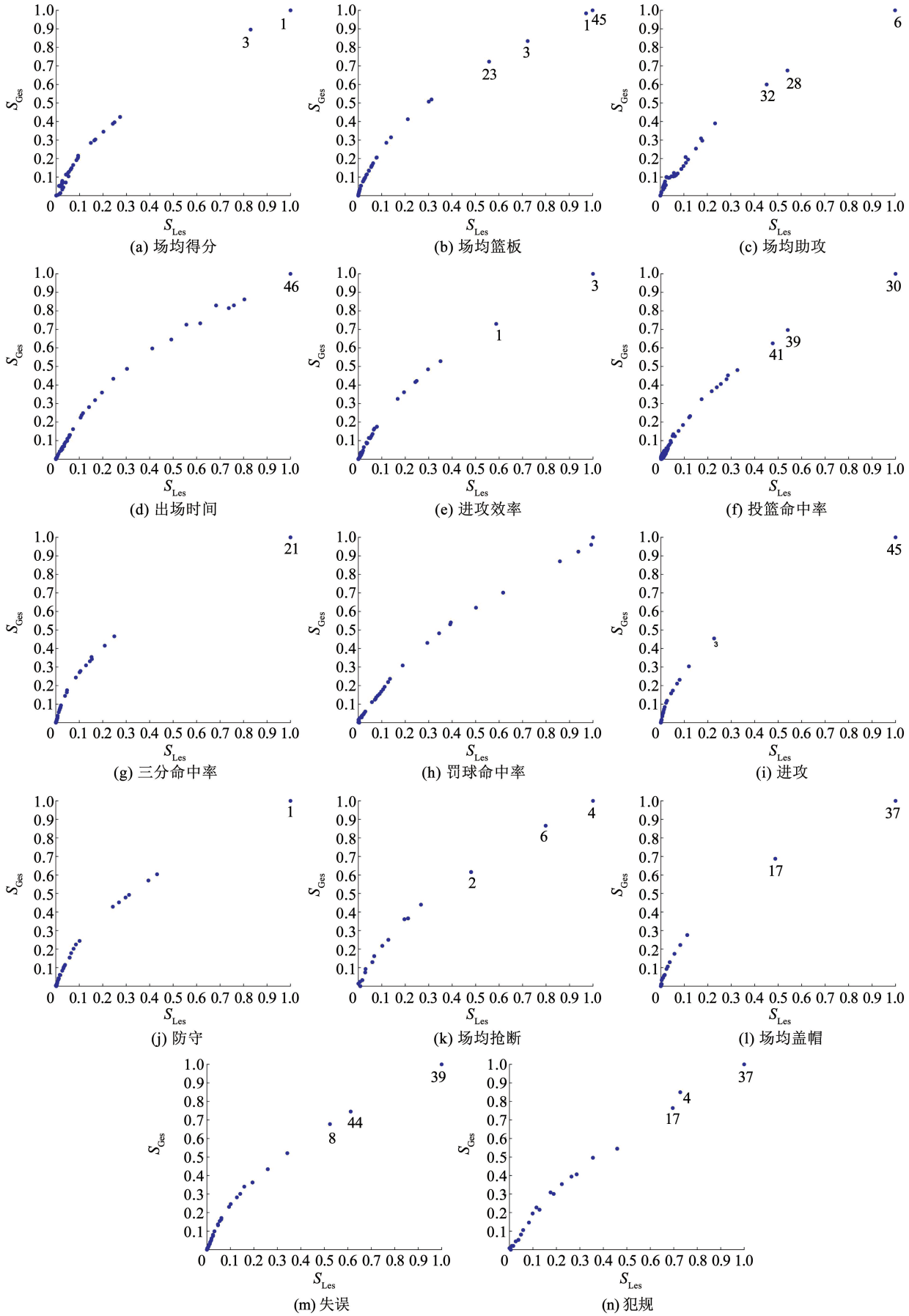


图 18　单个属性的 NBA 球员数据决策图

Fig. 18　Decision graph of single attribute in NBA player data

表 9 每个属性下的离群球员

Tab. 9 Outlier players under each attribute

离群点种类	场均得分	场均篮板	场均助攻	出场时间	进攻效率	投篮命中率	三分命中率	罚球命中率	进攻	防守	场均抢断	场均盖帽	失误	犯规
局部离群点	无	3、23	28、32	45	1	39、41	无	无	3	无	2、4、6	17	8、44	4、17
全局离群点	1、3	1、45	6	无	3	30	21	无	45	1	无	37	39	37

5.4 分析与讨论

1)由于传统 DPC 聚类算法计算局部密度时没有考虑数据的局部结构,所以当类簇之间具有不同的密度时,DPC 无法识别稀疏簇的聚类中心。本文利用 k 近邻方法和核密度估计方法计算数据点的局部密度,用此代替传统 DPC 聚类算法中根据截断距离计算的局部密度。该改进方法能提高 DPC 聚类算法在数据点有不同密度分布中的聚类效果,并提高聚类中心选取的准确度。

2)基于 k 近邻和聚类方法得到数据点的全局和局部异常值,并结合全局和局部异常值提高离群点的检测精度。

3)利用全局和局部异常值构建决策图,通过决策图观察数据点的种类,本文对全局和局部离群点进行解释。

4)常见的离群点检测方法受 k 影响较大,本文提出的基于改进 DPC 聚类算法的离群点检测方法受 k 影响较小。

6 结 论

1)利用 k 近邻和核密度估计方法计算数据点的局部密度,代替传统 DPC 聚类算法中根据截断距离计算的局部密度。该改进方法能提高 DPC 聚类算法在数据点有不同密度分布中的聚类效果,并提高聚类中心选取的准确度。接着,构建局部密度 - 相对距离决策图选取聚类中心并对数据点进行聚类。

2)通过 k 近邻方法计算数据点的全局异常值并计算簇的平均密度与数据点局部密度的比率得到局部异常值。将全局与局部异常值进行乘积得到最终的异常得分,选取异常值得分高的 Top-n 数据点作为离群点。通过人工数据集和 UCI 数据集实验发现,通过结合全局和局部异常值的方法能够提高离群点检测精度且受参数 k 影响较小。

3)本文提出一种构建全局 - 局部异常值决策图的离群点解释方法。通过人工数据集和 NBA 球员数据集的实验发现,全局离群点出现在决策图的右上方,而局部离群点稀疏分布在决策图的中间靠上部分。除此之外,还能观察出每个簇的正常数据点的密度分布情况。在面对未知数据集时,可以通过构建决策图的方式对数据集进行分析。

参考文献

[1] HAWKINS D M. Identification of outliers[M]. Dordrecht: Springer Netherlands, 1980. DOI: 10.1007/978-94-015-3994-4

[2] GAO Yongchang, GUAN Haowen, GONG Bin. CODM: an outlier detection method for medical insurance claims fraud[J]. International Journal of Computational Science and Engineering, 2019, 20(3): 404. DOI: 10.1504/ijcse.2019.103945

[3] HILAL W, GADSDEN S A, YAWNEY J. Financial fraud: a review of anomaly detection techniques and recent advances[J]. Expert Systems with Applications, 2022, 193: 116429. DOI: 10.1016/j.eswa.2021.116429

[4] ALHARBE N, ALI RAKROUKI M, ALJOHANI A. A healthcare quality assessment model based on outlier detection algorithm[J]. Processes, 2022, 10(6): 1199. DOI: 10.3390/pr10061199

[5] YANG Yun, FAN Chongjun, CHEN Liang, et al. IPMOD: an efficient outlier detection model for high-dimensional medical data streams[J]. Expert Systems with Applications, 2022, 191: 116212. DOI: 10.1016/j.eswa.2021.116212

[6] WU Huangjian, TANG Xiao, WANG Zifa, et al. Probabilistic automatic outlier detection for surface air quality measurements from the China national environmental monitoring network[J]. Advances in Atmospheric Sciences, 2018, 35(12): 1522. DOI: 10.1007/s00376-018-8067-9

[7] KANG S, KYUN KIM S. Outlier behavior detection for indoor environment based on t-SNE clustering[J]. Computers, Materials & Continua, 2021, 68(3): 3725. DOI: 10.32604/cmc.2021.016828

[8] LLANSÓ L, MOORE U, BOLANO-DIAZ C, et al. Expanding the muscle imaging spectrum in dysferlinopathy: description of an outlier population from the classical MRI pattern[J]. Neuromuscular Disorders, 2023, 33(4): 349. DOI: 10.1016/j.nmd.2023.02.007

[9] CHEN Zhaomin, YEO C K, LEE B S, et al. Evolutionary multi-objective optimization based ensemble autoencoders for image outlier detection[J]. Neurocomputing, 2018, 309: 192. DOI: 10.1016/j.neucom.2018.05.012

[10] RIBEIRO M, LAZZARETTI A E, LOPES H S. A study of deep convolutional auto-encoders for anomaly detection in videos[J]. Pattern Recognition Letters, 2018, 105: 13. DOI: 10.1016/j.patrec.2017.07.016

[11] LI Shifeng, LIU Chunxiao, YANG Yuqiang. Anomaly detection based on maximum a posteriori[J]. Pattern Recognition Letters, 2018, 107: 91. DOI: 10.1016/j.patrec.2017.09.001

[12] PANG Guansong, SHEN Chunhua, CAO Longbing, et al. Deep learning for anomaly detection: a review[J]. ACM Computing Surveys, 2021, 54(2): 38. DOI: 10.1145/3439950

[13] VILLA-PÉREZ M E, ÁLVAREZ-CARMONA M Á, LOYOLA-GONZÁLEZ O, et al. Semi-supervised anomaly detection algorithms: a comparative summary and future research directions[J]. Knowledge-

based Systems, 2021, 218: 106878. DOI: 10.1016/j.knosys.2021.106878

[14] ZHANG Ji. Advancements of outlier detection: a survey[J]. ICST Transactions on Scalable Information Systems, 2013, 13(1): e2. DOI: 10.4108/trans.sis.2013.01-03.e2

[15] 周玉, 朱文豪, 房倩, 等. 基于聚类的离群点检测方法研究综述[J]. 计算机工程与应用, 2021, 57(12): 37.
ZHOU Yu, ZHU Wenhao, FANG Qian, et al. Survey of outlier detection methods based on clustering[J]. Computer Engineering and Applications, 2021, 57(12): 37. DOI: 10.3778/j.issn.1002-8331.2102-0167

[16] ZHANG Zhongping, LI Sen, LIU Weixiong, et al. A new outlier detection algorithm based on fast density peak clustering outlier factor[J]. International Journal of Data Warehousing and Mining, 2023, 19(2): 1. DOI: 10.4018/ijdwm.316534

[17] RAMASWAMY S, RASTOGI R, SHIM K. Efficient algorithms for mining outliers from large data sets[J]. ACM SIGMOD Record, 2000, 29(2): 427. DOI: 10.1145/335191.335437

[18] ZHANG Ke, HUTTER M, JIN Huidong. A new local distance-based outlier detection approach for scattered real-world data[C]// Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin: Springer, 2009: 813. DOI: 10.1007/978-3-642-01307-2_84

[19] YANG Jiawei, RAHARDJA S, FRÄNTI P, et al. Mean-shift outlier detection and filtering[J]. Pattern Recognition, 2021, 115: 107874. DOI: 10.1016/j.patcog.2021.107874

[20] XIE Jiang, XIONG Zhongyang, DAI Qizhu, et al. A local-gravitation-based method for the detection of outliers and boundary points[J]. Knowledge-based Systems, 2020, 192: 105331. DOI: 10.1016/j.knosys.2019.105331

[21] BREUNIG M M, KRIEGEL H P, NG R T, et al. LOF: identifying density based local outliers[C]//Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data Dallas: ACM, 2000. DOI:10.1145/342009.335388

[22] TANG Jian, CHEN Zhixiang, FU A W C, et al. Enhancing effectiveness of outlier detections for low density patterns[M]// Advances in Knowledge Discovery and Data Mining. Berlin: Springer, 2002: 535. DOI: 10.1007/3-540-47887-6_53

[23] LATECKI L J, LAZAREVIC A, POKRAJAC D. Outlier detection with kernel density functions[C]//International Workshop on Machine Learning and Data Mining in Pattern Recognition. Berlin: Springer, 2007: 61. DOI: 10.1007/978-3-540-73499-4_6

[24] TANG Bo, HE Haibo. A local density-based approach for outlier detection[J]. Neurocomputing, 2017, 241: 171. DOI: 10.1016/j.neucom.2017.02.039

[25] 张忠平, 刘伟雄, 张玉停, 等. ERDOF: 基于相对熵权密度离群因子的离群点检测算法[J]. 通信学报, 2021,42(9):133
ZHANG Zhongping, LIU Weixiong, ZHANG Yuting. et al. ERDOF: outlier detection algorithm based on entropy weight distance and relative density outlier factor[J]. Journal on Communications, 2021, 42(9): 133. DOI: 10.11959/j.issn.1000-436x.2021152

[26] HE Zengyou, XU Xiaofei, DENG Shengchun. Discovering cluster-based local outliers[J]. Pattern Recognition Letters, 2003, 24(9/10): 1641. DOI: 10.1016/S0167-8655(03)00003-5

[27] AL-ZOUBI M B, AL-DAHOUD A, YAHYA A A. New outlier detection method based on fuzzy clustering[J]. WSEAS Transactions on Information Science and Applications, 2010, 7(5): 681

[28] 周玉, 朱文豪, 孙红玉. 一种基于目标函数的局部离群点检测方法[J]. 东北大学学报(自然科学版), 2022, 43(10): 1405
ZHOU Yu, ZHU Wenhao, SUN Hongyu. A local outlier detection method based on objective function[J]. Journal of Northeastern University (Natural Science), 2022, 43(10): 1405. DOI: 10.12068/j.issn.1005-3026.2022.10.006

[29] 张忠平,李森,刘伟雄,等. 基于快速密度峰值聚类离群因子的离群点检测算法[J]. 通信学报, 2022, 43(10): 186
ZHANG Zhongping, LI Sen, LIU Weixiong, et al. Outlier detection algorithm based on fast density peak clustering outlier factor[J]. Journal on Communications, 2022, 43(10): 186. DOI: 10.11959/j.issn.1000-436x.2022193

[30] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492. DOI: 10.1126/science.1242072

[31] DU Xusheng, YU Jiong, CHU Zheng, et al. Graph autoencoder-based unsupervised outlier detection[J]. Information Sciences, 2022, 608: 532. DOI: 10.1016/j.ins.2022.06.039

[32] LI Kangsheng, GAO Xin, JIA Xin, et al. Detection of local and clustered outliers based on the density-distance decision graph[J]. Engineering Applications of Artificial Intelligence, 2022, 110: 104719. DOI: 10.1016/j.engappai.2022.104719

[33] 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: ACM, 2008: 444. DOI: 10.1145/1401890.1401946

[34] LIU F T, TING Kaiming, ZHOU Zhihua. Isolation forest[C]// 2008 Eighth IEEE International Conference on Data Mining. Pisa: IEEE, 2008: 413. DOI: 10.1109/ICDM.2008.17

(编辑　张　红)