doi:10.3772/j.issn.1002-0470.2024.02.008

基于期望核密度离群因子的离群点检测算法①

张忠平②*** 孙光旭③* 姚春辰* 刘 硕* 齐文旭***

(*燕山大学信息科学与工程学院 秦皇岛 066004)

(**河北省计算机虚拟技术与系统集成重点实验室 秦皇岛 066004)

(***信息工程大学信息系统工程学院 郑州 450001)

摘 要 针对基于密度的离群点检测方法在不同分布的数据集上检测精度低的问题,提出了一种基于期望核密度离群因子的离群点检测算法。首先,引入k近邻和反向k近邻扩展邻域空间(ENS)代替传统的k邻域范围,更加全面地考虑数据对象的邻域信息;其次,在传统核密度估计(KDE)方法的基础上引入多元高斯函数,在扩展邻域空间内估计数据对象的密度,同时借鉴自适应核带宽的思想,更好地适应不同数据集的数据分布;然后,给出期望距离的概念,进一步区分局部离群点和位于低密度区域的正常点;最后,定义了期望核密度离群因子刻画数据对象离群程度。在人工数据集和真实数据集上对所提算法进行实验验证,并与部分传统算法进行对比,验证了所提算法的有效性。

关键词 数据挖掘;离群点;核密度估计(KDE);期望距离;期望核密度离群因子

离群点是指数据集中偏离大多数数据的少量数据对象,它们与正常数据对象存在明显的差异。离群点检测技术致力于消除噪音和干扰或发现潜在的、有价值的信息[1],是数据挖掘的一个重要研究方向。目前,离群点检测技术已经广泛应用于各大领域中,例如网络入侵检测[2-3]、金融欺诈检测[4-5]、工业损伤检测[6]、垃圾邮件检测[7]、医疗与公共卫生检测[8]等领域。

目前离群点检测大致可分为基于统计的[9-10]、基于距离的[11-12]、基于聚类的[13-14]和基于密度的[15-16]方法。

基于统计的方法采用统计学中的标准分布模型拟合数据,若某个数据点与假设的分布模型偏差较大,则视其为离群点。此类方法不适用于高维数据集,并且当数据集不服从任何标准分布或无法判断分布特征时,离群点检测效率将会大幅降低。

基于距离的方法通过计算数据对象之间的距离远近,将距离更远的数据对象标记为离群点,避免了数据分布假设。然而,此类方法只考虑了全局离群点,没有顾及到局部离群点。对此,Ramaswamy 等人[17]提出了一种基于距离的改进离群点检测算法——k近邻(k-nearest neighbors, KNN)算法。该算法首先对原始数据聚类,计算簇中样本点的k近邻距离的上下界,排除距离过小的簇,将剩余数据点中k近邻距离较大的点标记为离群点。文献[18]针对KNN算法对近邻参数 k 值敏感的问题,提出了一种基于自然最近邻的离群点检测算法,通过在不同数据集上自适应获取近邻参数,避免了人为设置。

基于聚类的方法将远离正常簇的离群聚类中的数据点以及不属于任何聚类的数据点视为离群点,此类方法通常会引入新的参数。文献[19]提出了一种基于累积全熵的子空间聚类离群点检测算法(sub-

① 国家自然科学基金(61972334),河北省创新能力提升计划(222567626H),中央引导地方科技发展资金项目(226Z1707G),四达铁路智能图像工件识别基金(No.x2021134)和秦皇岛城发健康产业发展有限公司绩效考核管理系统(x2022247)资助项目。

② 男,1972 年生,博士,教授;研究方向:大数据,数据挖掘,半结构化数据;E-mail:zpzhang@ysu.edu.cn。

③ 通信作者,E-mail:1957996858@qq.com。

(收稿日期:2023-06-20)

space outlier detection based on cumulative holoentropy, SODCH),该算法通过计算子空间的累积全熵值选取最优聚类子空间,提高检测效率。文献[20]提出了一种基于聚类离群因子和相互密度的离群点检测算法,算法根据相互邻居而非 k 邻域来计算数据的相互密度,通过构造决策图完成聚类,进而识别出离群点。

基于密度的方法是当下离群点检测领域的研究热点。传统的基于密度的方法大多将密度视作距离的倒数,通过数据对象间的距离来计算局部密度,离群点与位于密集区域的正常对象相比密度更低。局部离群因子(local outlier factor, LOF)算法[21]通过计算每个数据对象的局部离群因子来检测数据集中的离群点,作为迄今为止最为经典的离群点检测算法,仍存在精确率低、参数设置敏感等问题。Huang 等人[22]针对 LOF 近邻参数 k 选取困难的问题,提出了无参数的基于自然邻居的离群点检测算法——自然离群因子(natural outlier factor,NOF)算法。该算法合并 k 邻域和反 k 邻域,自适应获取近邻参数 k,通过定义离群因子 NOF 检测离群点。Li 等人[23]提出了一种基于密度-距离决策图的离群点检测算法,将传统核密度估计(kernel density estimation,KDE)与局部可达距离相结合检测局部离群点,根据密度提升距离的度量标准检测全局离群点,通过密度比和密度提升距离生成决策图,同时检测出局部、全局和聚类离群点。

通过将核密度估计应用于离群点检测算法中,Wahid 等人[24]提出了一种基于相对核密度的离群点检测(relative kernel-density outlier factor,KDOF)算法。该算法使用核密度估计计算数据对象的局部密度,然后计算密度波动及其近邻点的密度差的平方和,最后通过比较其密度波动与近邻点的平均密度波动来描述数据对象的离群程度。基于此,Wahid 等人[25]又提出了一种基于自然邻居的离群点检测(natural neighbor outlier detection, NaNOD)算法,根据自然邻居自适应获取近邻参数 k,并采用加权核密度估计计算数据对象的密度,采用高斯核函数保证数据点之间的平滑度,使用自适应核带宽的思想适应数据点之间的密度差异,进一步区分正常点和离群点。上述 2 种算法在很多场景下都具有良好的性能,但存在维度灾难的问题,导致算法运行时间过长,离群点检测效率显著下降。

本文针对基于密度的离群点检测方法在不同分布的数据集上检测精度低的问题,提出一种基于期望核密度离群因子的离群点检测(outlier detection algorithm based on expected kernel density outlier factor, EKDOF)算法。首先,算法将 k 近邻和反向 k 近邻合并为邻域空间,充分考虑数据对象的局部信息;其次,将核密度估计与多元高斯函数相结合估计数据对象的局部密度,同时根据数据点的 k 邻域平均距离自适应地选取核带宽,给出了期望距离的概念,进一步区分局部离群点和低密度区域的正常点;最后,利用期望距离与核密度估计的比值定义期望核密度离群因子来刻画数据对象的离群程度。

1 相关工作

下面简要介绍相互 k 近邻数搜索算法。其中,D 表示数据集,p、q、o 为数据集 D 中的数据点,dist(p,q)表示点 p、q 之间的欧式距离,k 为正整数。

1.1 扩展邻域空间

定义 1 k 距离 $d_k(p)$。$d_k(p)$ 是指在给定的 k 值下,点 p 到点 m 之间的欧氏距离。其中,数据点 m 满足:至少存在 k 个数据点 $m' \in D\backslash\{m'\}$,满足 $d(p,m') \leqslant d(p,m)$。

定义 2 数据点 x_i 的 k 近邻 $kNN(x_i)$[22]。$kNN(x_i)$ 是指在数据集 $D = \{x_1,x_2,\cdots,x_n\}$ 中,到数据点 x_i 的距离不大于数据点 x_i 的 k 距离的点的集合,定义为式(1)。

$$kNN(x_i) = \{x_j \in D \mid d(x_i,x_j) \leqslant d_k(x_i)\} \tag{1}$$

定义 3 数据点 x_i 的反向近 $RNN(x_i)$[22]。$RNN(x_i)$ 是指在数据集 $D = \{x_1,x_2,\cdots,x_n\}$ 中,将数据点 x_i 看作 k 最近邻居的数据点 x_j 所构成的点的集合,定义为式(2)。

$$RNN(x_i) = \{x_j \in D \mid x_i \in kNN(x_j)\} \tag{2}$$

定义 4 扩展邻域空间(extended neighborhood space, ENS) $ENS(x_i)$[26]。$ENS(x_i)$ 是指数据点 x_i

的 k 近邻 $kNN(x_i)$ 和反向近邻 $RNN(x_i)$ 的并集所组成的集合,定义为式(3)。

$$ENS(x_i) = kNN(x_i) \cup RNN(x_i) \tag{3}$$

1.2 核密度估计

在数据分布未知的情况下,使用核密度估计可以估计出每个数据对象本身的局部密度,更好地反映出数据点本身与其扩展邻域空间内其他数据点之间的密度差异。

定义 5 核密度估计(KDE)[27]。KDE 是指通过估计样本集中样本的概率密度函数,得出样本集的总体分布情况,定义为式(4)。

$$\rho(x_i) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{h(x_j)^d}K\left(\frac{\| x_i - x_j \|}{h(x_j)}\right) \tag{4}$$

其中,n 表示数据集 D 中数据点的个数;$h(x_j)$ 表示在数据点 x_j 上的带宽,也称为平滑函数;d 表示数据点的维度;$\| x_i - x_j \|$ 表示数据点 x_i 到数据点 x_j 的欧式距离。$K()$表示核函数,具有期望值为 0、非负性和归一性的特征,满足以下条件:

$$\int K(x)\mathrm{d}x = 1 \tag{5}$$

$$\int xK(x)\mathrm{d}x = 0 \tag{6}$$

$$\int x^2K(x)\mathrm{d}x > 0 \tag{7}$$

在传统计算核密度的方法中,通常根据经验值将带宽设置为固定带宽,难以体现出不同数据点之间的局部密度差异。局部密度因子(local density factor, LDF)算法[28],提出一种自适应获取核带宽的方法,以适应在不同数据集下数据点局部密度的差异,定义为式(8)。

$$h(x_i) = h \cdot d_k(x_i) \tag{8}$$

其中,$d_k(x_i)$ 表示数据点 x_i 到第 k 个近邻点的距离,但该方法仍然需要用户事先指定 h 的取值。

2 EKDOF 算法

2.1 EKDOF 算法相关定义

定义 6 自适应核带宽 $h(xj)$。$h(xj)$ 是指在给定近邻参数 k 值时,2 个数据对象的度量参数乘积的拟合值,定义为式(9)。

$$h(x_j) = (m_i \cdot m_j)^{\frac{1}{2}} \tag{9}$$

其中,m_i、m_j 分别表示数据点 x_i、x_j 的度量参数,且数据点 x_j 在数据点 x_i 的扩展邻域空间内,即 $x_j \in ENS(x_i)$。自适应核带宽能够在估计密度时根据数据对象所处区域疏密程度的不同而发生改变。

定义 7 度量参数 m_i。m_i 是指数据点的 k 邻域平均距离,定义为式(10)。

$$m_i = \frac{\sum_{x_l \in kNN(x_i)} d(x_i, x_l)}{| kNN(x_i) |} \tag{10}$$

其中,$kNN(x_i)$ 表示数据点 x_i 的 k 近邻点的集合,$| kNN(x_i) |$ 表示数据点 x_i 的 k 邻域范围内数据点的个数,$d(x_i, x_l)$ 表示数据点 x_i 到数据点 x_l 的欧式距离。

定义 8 多元高斯函数 $K()$。$K()$ 是核密度估计的底层函数,定义为式(11)。

$$K(x) = \frac{1}{(2\pi)^d}\exp(-\frac{\| \boldsymbol{x} \|^2}{2}) \tag{11}$$

其中,d 为数据对象的维度,$\| \boldsymbol{x} \|$ 表示向量 $\boldsymbol{x}$ 的模。

定义 9 核密度估计 $den_{\mathrm{ENS}}(x_i)$。$den_{\mathrm{ENS}}(x_i)$ 是指在 x_i 的扩展邻域空间内,结合多元高斯函数和传统核密度估计得到的密度,定义为式(12)。

$$den_{\mathrm{ENS}}(x_i) = \frac{1}{| ENS(x_i) |}\sum_{x_j \in \mathrm{ENS}(x_i)} \frac{1}{(2\pi)^d h(x_j)^d}\exp\left(-\frac{\| x_i - x_j \|^2}{2h(x_j)^2}\right) \tag{12}$$

其中,$| \mathrm{ENS}(x_i) |$ 表示数据点 x_i 的扩展邻域空间内数据对象的个数。综合式(9)和(12),得到 $den_{\mathrm{ENS}}(x_i)$ 的最终计算式定义为式(13)。

$$den_{\mathrm{ENS}}(x_i) = \frac{1}{| \mathrm{ENS}(x_i) |}\sum_{x_j \in \mathrm{ENS}(x_i)} \frac{1}{(2\pi)^d (m_i m_j)^{\frac{d}{2}}}\exp\left(-\frac{\| x_i - x_j \|^2}{2h(x_j)^2}\right) \tag{13}$$

定义 10 期望 k 距离 Ek_dist(expected k-distance)。Ek_dist 是指所有数据点到其各自的第 k 个近邻点的距离平均值,定义为式(14)。

$$Ek\_\mathrm{dist} = \frac{1}{n}\sum_{i=1}^{n} d(x_i, N_{k\text{-th}}) \tag{14}$$

其中,$N_{k\text{-th}}$ 表示第 k 个近邻点,x_i 为数据集 D 中的

数据点，n 表示数据集 D 中样本个数。期望 k 距离是数据集中的每个数据点到第 k 个近邻点的期望值，反映了在不同 k 值下，数据点与近邻点之间的距离偏差。

定义 11　期望 k 距离差（difference of expected k-distance）$dif_{Ek\_\mathrm{dist}}(x_i)$。$dif_{Ek\_\mathrm{dist}}(x_i)$ 指数据点 x_i 到第 k 近邻的距离与期望 k 距离的差，定义为式(15)。

$$dif_{Ek\_\mathrm{dist}}(x_i) = d(x_i, N_{k\text{-th}}) - Ek\_\mathrm{dist} \tag{15}$$

其中，$d(x_i, N_{k\text{-th}})$ 表示数据点 x_i 到第 k 个近邻点的欧式距离。可以看出，期望 k 距离差 $dif_{Ek\_\mathrm{dist}}(x_i)$ 的结果可取到正值、负值或者零。如果数据点到第 k 个近邻点的距离大于期望 k 距离，则 $dif_{Ek\_\mathrm{dist}}(x_i)$ 的值为正，否则为负；如果数据点到第 k 个近邻点的距离刚好等于期望 k 距离，则 $dif_{Ek\_\mathrm{dist}}(x_i)$ 的值等于 0。期望 k 距离差的正值越大，代表数据点本身与邻居点之间越疏远，该点越有可能是离群点；期望 k 距离差的负值结果越小，代表数据点本身与邻居点之间越紧密，该点越有可能是正常点。

图 1 展示了一个二维数据集部分数据点的分布情况。O_1、O_2、O_3 分别表示点 p 的 3 个近邻点；实线表示数据集中所有数据点到第 k 个近邻点的平均距离，即期望 k 距离，分别用 $E1\_\mathrm{dist}$、$E2\_\mathrm{dist}$ 和 $E3\_\mathrm{dist}$ 表示；虚线表示数据点 p 到第 k 个近邻点的距离与期望 k 距离的差，分别用 $dif_{E1\_\mathrm{dist}}(p)$、$dif_{E2\_\mathrm{dist}}(p)$、$dif_{E3\_\mathrm{dist}}(p)$ 表示。从图 1 可以看出，数据点 p 到第 1 个近邻点 O_1 的距离大于 $E1\_\mathrm{dist}$，因此 $dif_{E1\_\mathrm{dist}}(p)$ 的值为正；数据点p到第2个近邻点O_2的距离小于 $E2\_\mathrm{dist}$，因此 $dif_{E2\_\mathrm{dist}}(p)$ 的值为负；数据点 p 到第 3 个近邻点 O_3 的距离大于 $E3\_\mathrm{dist}$，因此 $dif_{E3\_\mathrm{dist}}(p)$ 的值为正。

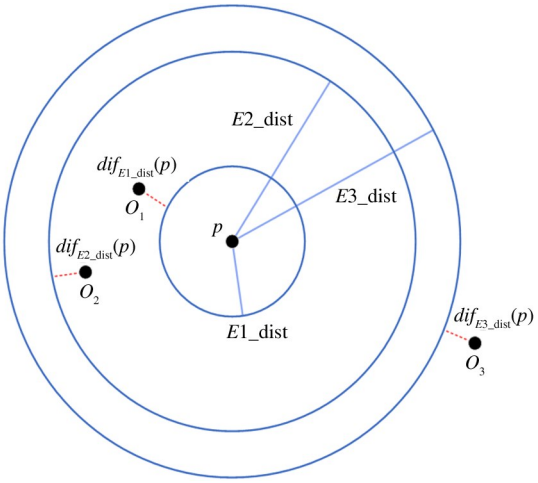


图 1　$k=3$ 时，期望 k 距离和期望 k 距离差的简单示例

定义 12　期望距离（expected distance）$E\mathrm{dist}(x_i)$。$E\mathrm{dist}(x_i)$是指在给定 k 值下，数据点 x_i 的期望 k 距离差的和，定义为式(16)。

$$E\mathrm{dist}(x_i) = \sum_{k=1}^{s} dif_{Ek\_\mathrm{dist}}(x_i) \tag{16}$$

其中，s 表示近邻参数 k 的值。位于密集区域的正常点与其邻域范围内的数据点之间的相互距离较小，则期望距离的值也会较小；位于稀疏区域的离群点与其邻域范围内的数据点之间的相互距离较大，则期望距离的值也会较大，从而可以进一步区分离群点和位于低密度区域的正常点。

定义 13　期望核密度离群因子 $EKDOF(x_i)$。$EKDOF(x_i)$ 是指数据点 x_i 的期望距离与核密度估计的比值，定义为式(17)。

$$EKDOF(x_i) = \frac{E\mathrm{dist}(x_i)}{den_{\mathrm{ENS}}(x_i)} \tag{17}$$

其中，$E\mathrm{dist}(x_i)$ 表示数据对象的期望距离，$den_{\mathrm{ENS}}(x_i)$ 表示数据对象的核密度估计，两者的比值表明了数据对象的离群程度。由于离群点常常偏离正常点，从而离群点到其第 k 个近邻点的距离总是大于期望 k 距离 Ek_dist，其期望 k 距离差 $dif_{Ek\_\mathrm{dist}}(x_i)$ 也会更大且总是取到正值，因此离群点的期望距离 $E\mathrm{dist}(x_i)$ 要远大于正常点，又由于离群点总是位于低密度区域，其核密度 $den_{\mathrm{ENS}}(x_i)$ 一般较小。因为期望距离 $E\mathrm{dist}(x_i)$ 更大、核密度 $den_{\mathrm{ENS}}(x_i)$ 更小，所以离群因子 $EKDOF(x_i)$ 的值也会更大，该点是离群点的概率也就越大。

2.2　EKDOF 算法思想

EKDOF 算法思想如下。首先，扩大数据对象的邻域范围，引入反向 k 近邻关系，将 k 近邻和反向 k 近邻合并作为数据对象的扩展邻域空间，完善了仅考虑 k 邻域范围内数据点分布情况的不足；采用自适应核带宽的思想计算核密度，在带宽函数中引入一种度量参数，根据不同数据点之间度量参数的大小来判断数据点所处区域的密集或稀疏程度。核带宽会随着数据点的变化而变化，而不是一个固定值。当数据点位于密集区域时，数据点之间的 k 邻域平

均距离小,度量参数的值变小,核带宽也就越小;当数据点位于稀疏区域时,数据点之间的 k 邻域平均距离大,度量参数的值变大,核带宽也就随之变大。将高斯核函数应用到核密度估计中,在扩展的邻域空间内计算数据点的局部密度。由于基于密度的方法存在低密度模式的问题,位于低密度区域内的正常点和位于密集区域边界的局部离群点的密度大小较为接近,仅凭密度难以区分,因此本文提出期望距离的概念。位于低密度区域内的数据点的期望距离相对较小,而位于密集区域边界的局部离群点的期望距离相对较大,可以进一步区分正常点和离群点。本文将高斯核密度估计与期望距离相结合构造离群因子,通过比较离群因子值的大小检测离群点。

2.3 EKDOF 算法描述

根据相关定义和算法思想,本节提出了基于期望核密度离群因子的离群点检测算法 EKDOF,算法描述如算法 1 所示。

算法 1 EKDOF 离群点检测算法

输入:数据集 D,近邻参数 k;

输出:离群点 S_{outlier}。

EKDOF(D,k)

BEGIN

1. 初始化 $kNN(x_i)=\varnothing$、$RNN(x_i)=\varnothing$、$EKDOF(x_i)=\varnothing$、$S_{\text{outlier}}=\varnothing$;
2. FOR 每个数据点 $x_i \in D$:
3. 计算数据集中每个数据点 x_i 的 k 近邻点 $kNN(x_i)$ 和反向 k 近邻点 $RNN(x_i)$;
4. 根据式(3)得到数据点 x_i 的扩展邻域空间 $ENS(x_i)$;
5. FOR 每个数据点 $x_j \in ENS(x_i)$:
6. 根据式(9)、(10)计算 x_i、x_j 的核带宽 $h(x_j)$;
7. END FOR
8. 根据式(13)计算 x_i 的核密度估计 $den_{\text{ENS}}(x_i)$;
9. 根据式(14)、(15)计算 x_i 的期望 k 距离差 $dif_{Ek\_\text{dist}}(x_i)$;
10. 根据式(16)计算 x_i 的期望距离 $E\text{dist}(x_i)$;
11. 根据式(17)计算 x_i 的期望核密度 $EKDOF(x_i)$;
12. END FOR
13. 根据 $EKDOF(x_i)$ 的值,对所有数据点降序排序;
14. 输出前 n 个数据点作为离群点集合 S_{outlier}。

END

首先,将数据点的 k 近邻与反向 k 近邻合并形成扩展邻域空间,在扩展的邻域空间内,通过将传统核密度估计与多元高斯函数相结合估计每个数据点的密度,同时引入度量参数的概念,根据数据对象的 k 邻域平均距离自适应获取数据对象之间的核带宽;其次,根据期望 k 距离和期望 k 距离差的定义计算每个数据对象的期望距离;最后,用数据对象的期望距离与核密度估计的比值构造离群因子,根据离群因子值进行降序排序,将离群因子值较大的前 n 个点输出即为离群点。

算法 1 中,$EKDOF(x_i)$ 中存放的分别是每个数据点 $x_i \in D$ 的期望核密度离群因子的值,算法将 $EKDOF(x_i)$ 值较大的前 n 个点输出作为离群点,S_{outlier}中存放的是算法最终检测到的所有离群点。

2.4 EKDOF 算法分析

2.4.1 正确性分析

在 EKDOF 算法中,首先将数据对象的 k 近邻和反向 k 近邻合并生成扩展邻域空间,充分考虑数据点的邻域信息;将传统核密度估计与多元高斯函数相结合估计数据对象的密度,同时引入自适应核带宽的思想,能够在不同分布的数据集下根据数据对象邻域范围的密集或稀疏程度自动进行调整,平滑了正常点之间的密度差异;通过分析发现,仅凭密度难以辨别出位于低密度区域的正常点和位于密集簇边界区域的局部离群点,因此本算法提出期望距离的概念,能够进一步区分正常点和局部离群点。EKDOF 算法采用数据对象的期望距离与核密度估计的比值构造离群因子,离群因子值越大,该点是离群点的可能性就越大。以上分析说明了所提出的基于期望核密度离群因子的离群点检测算法在原理上是正确的。

2.4.2 时间复杂度分析

EKDOF 算法的时间复杂度主要分为 2 部分:(1)在构造扩展邻域空间的过程中,需要计算每个数据点的 k 近邻和反向 k 近邻,时间复杂度为 $O(n\cdot\log n)$,其中 n 表示数据集中数据点的个数;(2)计算数据对象的期望核密度离群因子 $EKDOF(x_i)$,时间复杂度为 $O(n)$。综合上述分析,所提算法的时间复杂度为 $O(n\cdot\log n)+O(n)$ 两部分之和,进而得出 EKDOF 算法最终的时间复杂度为 $O(n\cdot\log n)$。

3　实验与分析

3.1　实验环境配置

表 1 给出了验证本文算法性能所采用的实验环境，主要包括软硬件环境以及各配置所对应的参数。

表 1　实验环境

软硬件环境	参数
CPU	Intel Core i5-8300H 2.30 GHz
内存	8.0 GB
硬盘	1.0 TB
操作系统	64 bit Windows 10
开发环境	Pycharm
编译环境	Python 3.10
可视化工具	Python 3.10

3.2　实验评价指标

本文采用的实验评价指标为精确率 Pr、AUC（area under curve）值和离群点发现曲线。

精确率 Pr 是指算法实际检测到的离群点数量与离群点总数之比，定义为式(18)。

$$Pr = \frac{TP}{TP + FP} \tag{18}$$

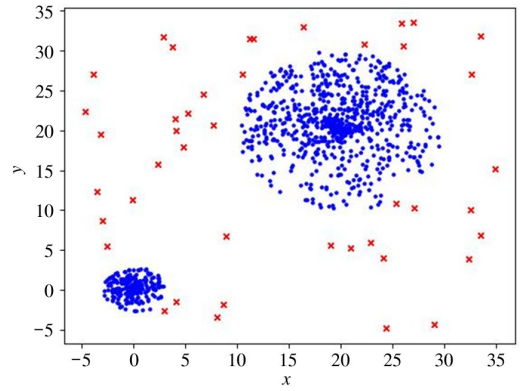
其中，FP 表示算法把数据集中的正常点当作离群点进行输出的数量，TP 表示算法最终检测到的真实离群点的个数。精确率 Pr 的取值在 0～1 之间，Pr 的值越接近于 1，代表算法检测出的真实离群点的数量越多，算法的性能就越好。

AUC 值是一个介于 0 和 1 之间的数，AUC 的值越接近于 1，代表算法越接近于离群点检测的最佳水平，能把更多的离群点排在正常点之前进行输出，算法的性能就越好。

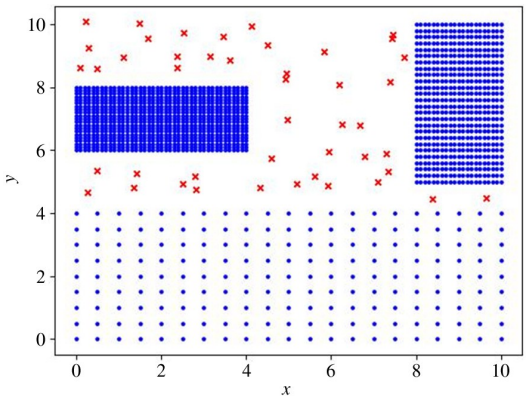
离群点发现曲线[29]反映的是算法实际检测到的离群点个数随用户查询个数的变化趋势。离群点发现曲线的横坐标是用户想要查询的离群点的个数，纵坐标是算法真正检测到的离群点的个数。离群点发现曲线的斜率越接近于 1，代表算法越能满足用户的查询需求，算法的性能就越好。

3.3　人工数据集实验与分析

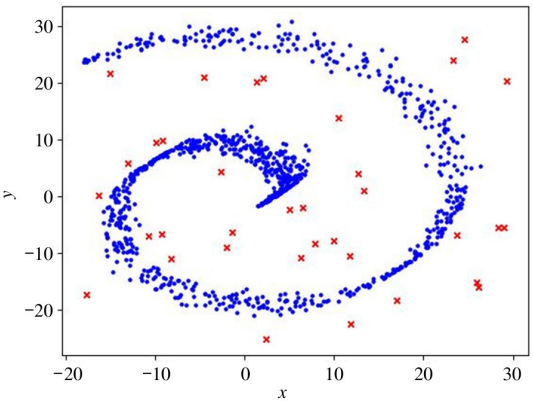
本文使用图 2 所示的二维人工数据集 DS5～DS8



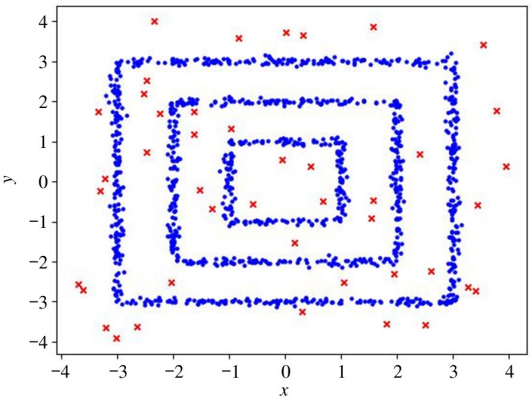
（a）DS5 数据集分布



（b）DS6 数据集分布



（c）DS7 数据集分布



（d）DS8 数据集分布

图 2　EKDOF 算法人工数据集的数据分布图

进行对比实验,其中“×”代表离群点,其余点代表正常点。DS5 ~ DS8 的数据特征如表 2 所示。

表 2　人工数据集的数据特征

数据集	样本数	离群点个数	离群点比例/%
DS5	1 043	43	4.1
DS6	1 641	45	2.7
DS7	1 031	35	3.3
DS8	1 256	43	3.4

结合图 2 和表 2 可以看出,本算法选取的人工数据集的数据分布复杂程度不同,既包括离群点和正常点分布较为明显的数据集,也包括离群点交错地分布在正常点所构成的聚类簇的周围或边界区域的数据集。因此,在这些数据集上进行实验可以有效地验证 EKDOF 算法的性能,同时有助于与其他对比算法作出区分。

表 3 和图 3 展示了在 4 个人工数据集下,EKDOF 算法和其他 4 种对比算法在精确率上的实验结果。

表 3　不同算法在人工数据集上的精确率

数据集	INFLO	NOF	LOLED	LDF	EKDOF
DS5	0.93	0.69	0.79	0.90	**0.95**
DS6	0.75	0.44	0.73	0.82	**0.88**
DS7	0.82	0.37	0.82	0.82	**0.91**
DS8	0.93	0.79	0.86	0.93	**0.97**

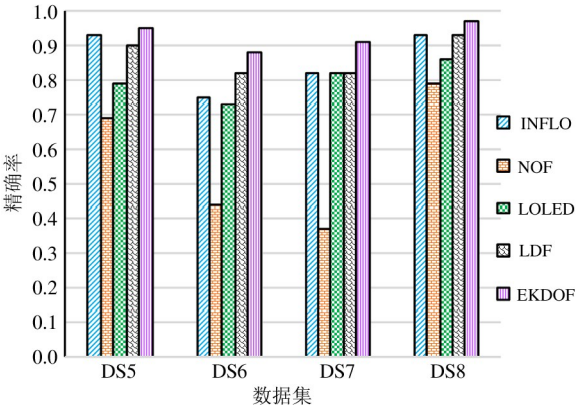


图 3　不同算法在人工数据集上的精确率对比

结合表 3 和图 3 可以看出,EKDOF 算法在人工数据集上的精确率均为最优,且在 DS8 数据集上,EKDOF 算法的精确率为 0.97,在所有人工数据集中达到了最高。在所有数据集中 EKDOF 算法的精确率均高于同样使用了核密度估计方法的 LOLED 算法和 LDF 算法,特别是在 DS5 数据集上 EKDOF 算法的精确率为 0.95,而 LOLED 算法的精确率为 0.79。这是因为 EKDOF 算法同时将 k 邻域和反向 k 邻域作为邻域空间,更加全面地考虑数据对象的局部信息,使用期望距离能够进一步区分离群点和位于低密度区域的正常点,因此提高了本文算法的检测精度。在 DS6 数据集上,每种算法的精确率均有所下降,EKDOF 算法的精确率为 0.88,仍高于对比算法。在 4 个人工数据集上,EKDOF 算法的精确率均优于同样使用了扩展邻域空间的 INFLO 算法和 NOF 算法,特别是在 DS7 数据集上,EKDOF 算法的精确率为 0.91,高于 NOF 算法 0.54。综上所述,EKDOF 算法在所有数据集上的精确率都是最高的,从而验证了本文所提出的基于期望核密度离群因子的离群点算法在精确率上的有效性。

表 4 和图 4 展示了 4 个人工数据集上,EKDOF 算法和其他 4 种对比算法在 AUC 值上的实验结果。

表 4　不同算法在人工数据集上的 AUC 值

数据集	INFLO	NOF	LOLED	LDF	EKDOF
DS5	0.96	0.82	0.82	0.89	**0.97**
DS6	0.88	0.75	0.87	0.91	**0.97**
DS7	0.98	0.30	0.81	0.79	**1.00**
DS8	0.93	0.54	0.96	0.90	**1.00**

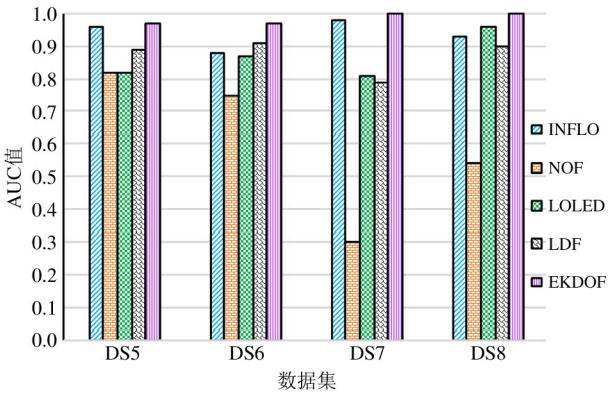
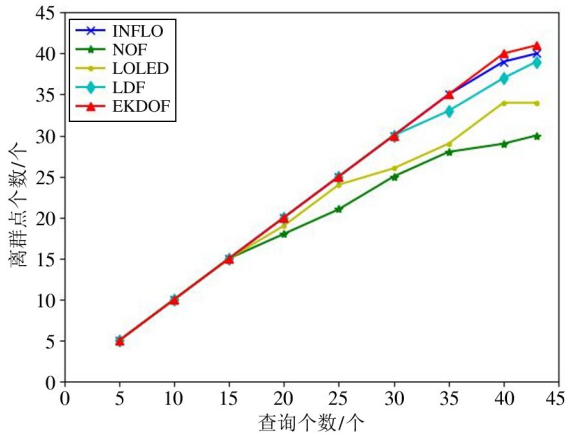


图 4　不同算法在人工数据集上的 AUC 值对比

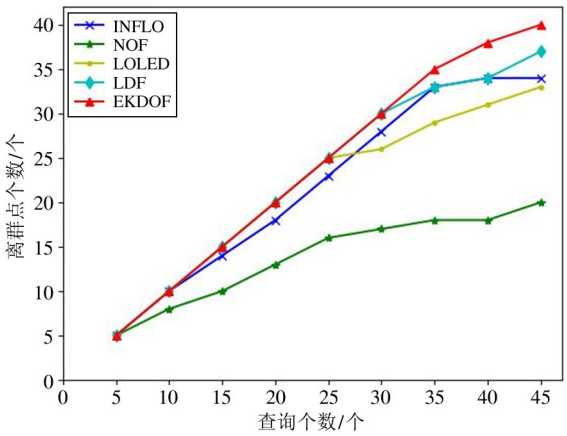
结合表 4 和图 4 可以看出,EKDOF 算法在人工数据集上的 AUC 值均为最优,其中在 DS7 和 DS8 数据集上,其 AUC 值均达到了 1.00。在 DS5 和 DS6 数据集上,EKDOF 算法的 AUC 值均为 0.97,除 NOF

算法在 DS6 数据集上的 AUC 值为 0.75 之外，其他算法在这 2 个数据集上的 AUC 值均保持在 0.80 以上。在 DS7 数据集上，LOLED 算法、LDF 算法和 NOF 算法的 AUC 值均有所下降，其中 NOF 算法的 AUC 值仅为 0.30，这是因为 DS7 数据集分布较为复杂，通过密度难以区分出局部离群点和低密度区域的正常点，难以检测出更多的离群点，因此造成算法 AUC 值减小。综上，EKDOF 算法在 4 个人工数据集上的 AUC 值都是最高的，证明本文提出的基于期望核密度离群因子的离群点检测算法在 AUC 值上的有效性。

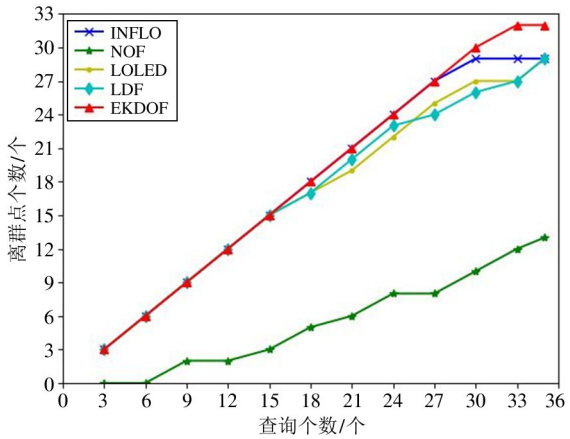
图 5 展示了 4 个人工数据集上，EKDOF 算法和 4 种对比算法的离群点发现曲线的实验结果。



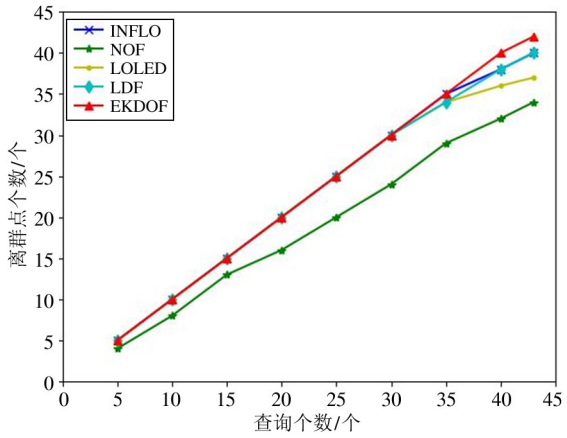
（a）DS5 离群点发现曲线



（b）DS6 离群点发现曲线



（c）DS7 离群点发现曲线



（d）DS8 离群点发现曲线

图 5　不同算法在人工数据集上的离群点发现曲线对比

观察图 5(a) ~ (d)的曲线变化趋势可以看出，当查询条件相同时，EKDOF 算法能够查询到更多的离群点反馈给用户，表明 EKDOF 算法相比于其他算法具有更优秀的检测能力。NOF 算法在 DS8 数据集上的表现和其他算法较为相似，但在 DS6 和 DS7 数据集上的检测效果与其他算法相比相差较多。在 DS8 数据集上，EKDOF 算法的离群点发现曲线的斜率最接近于 1，当用户的查询数量从 5 依次递增至 40 时，EKDOF 算法查询到的离群点个数与用户设置的查询个数始终相等，说明 EKDOF 算法检测到的前 40 个数据点都是离群点。综上所述，EKDOF 算法的离群点发现曲线在 4 个人工数据集上的表现都是最优的，从而验证了本文所提出的 EKDOF 算法在离群点检测方面的良好性能。

3.4　真实数据集实验与分析

4 个真实数据集的数据特征如表 5 所示，它们

均来自于 UCI 真实数据集[30]。由表 5 可知,EKDOF 算法所选取的真实数据集的样本数量从 129 ~ 1 641,数据维度从 7 ~ 17,离群点数量从 10 ~ 25,离群点占比从 1.2% ~ 14.8%。由此可知,本算法选取的真实数据集的数据分布较为复杂,有包含几百个数据点的小规模数据集,也有包含上千个数据点的大规模数据集。因此,在这些数据集上进行实验可以有效地对比各离群点检测算法的性能。

表 5　真实数据集的数据特征

数据集	样本数	维度	离群点个数	离群点比例/%
PenDigits	1 641	17	20	1.2
Ecoli	168	7	25	14.8
WBC	223	9	10	4.4
wine	129	13	10	7.7

表 6 和图 6 展示了在 4 个真实数据集下,EKDOF 算法和其他 4 种对比算法在精确率上的实验结果。

表 6　不同算法在真实数据集上的精确率

数据集	INFLO	NOF	LOLED	LDF	EKDOF
PenDigits	0.90	0.20	0.25	0.20	**0.95**
Ecoli	**0.84**	0.48	0.80	0.76	**0.84**
WBC	0.60	0	0.20	0.10	**0.90**
wine	0.70	0.10	0.70	0.50	**0.90**

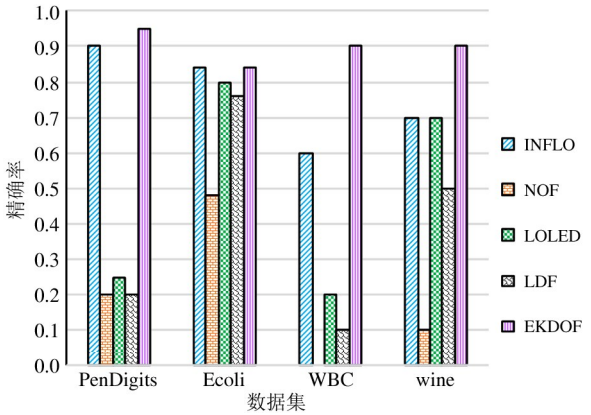


图 6　不同算法在真实数据集上的精确率对比

结合表 6 和图 6 可以看出,除了在 Ecoli 数据集上本文算法和 INFLO 算法的精确率相等以外,在其他 3 个数据 EKDOF 算法的精确率都要比另外几种算法好,特别是在 PenDigits 数据集上达到了 0.95。INFLO 算法在对比算法中的表现相对较好,而另外 3 种对比算法的精确率表现较差且波动幅度较大。NOF 算法的稳定性最差,虽然在 Ecoli 数据集上有 0.48 的检测精度,但在 WBC 数据集上根本没有检测出离群点,而本文算法的精确率仍然可以达到 0.90。综上所述,EKDOF 算法在每个真实数据集上都能检测出更多的离群点且稳定性也是最好的,从而验证了本文所提出基于期望核密度离群因子的离群点检测算法在精确率上的有效性。

表 7 和图 7 展示了在 4 个真实数据集下,EKDOF 算法和其他 4 种对比算法在 AUC 值上的实验结果。

表 7　不同算法在真实数据集上的 AUC 值

数据集	INFLO	NOF	LOLED	LDF	EKDOF
PenDigits	**1.00**	0.60	0.90	0.82	**1.00**
Ecoli	0.98	0.48	0.91	0.90	**1.00**
WBC	0.41	0.00	0.50	0.77	**1.00**
wine	0.80	0.66	0.52	0.76	**0.88**

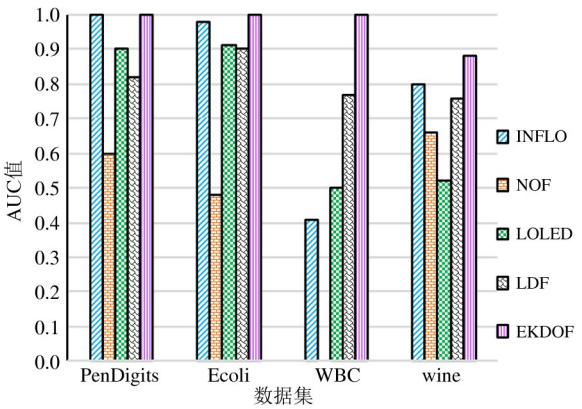
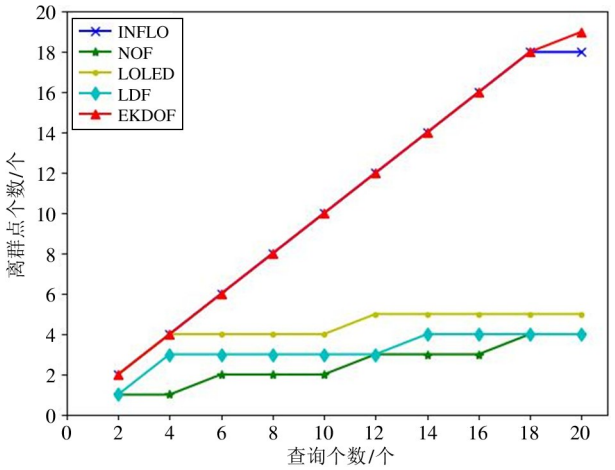


图 7　不同算法在真实数据集上的 AUC 值对比

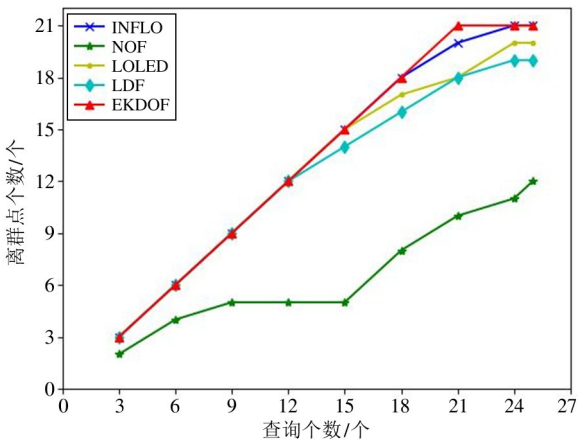
结合表 7 和图 7 可以看出, EKDOF 算法在每个真实数据集上的 AUC 值都有着不错的实验结果,在 PenDigits、Ecoli 和 WBC 这 3 个数据集上,EKDOF 算法的 AUC 值均为 1.00,说明 EKDOF 算法能把检测出来的所有离群点排在正常点之前进行输出。在 PenDigits 数据集上,EKDOF 算法和 INFLO 算法的 AUC 值相同。NOF 算法在所有真实数据集上的 AUC 值和稳定性表现仍是最差的,由于 NOF 算法在 WBC 数据集上未检测出离群点,其 AUC 值为 0.00。

INFLO 算法的稳定性也一般,在 WBC 数据集上的 AUC 值仅有 0.41,但在 PenDigits 数据集上的 AUC 值能达到 1.00。LDF 算法在 4 个对比算法中的表现最为平稳,其 AUC 值一直保持在 0.75 以上。综合上述的分析,EKDOF 算法在每个真实数据集上的 AUC 值都是最高的,从而验证了本文所提出的基于期望核密度离群因子的离群点检测算法在 AUC 值上的有效性。

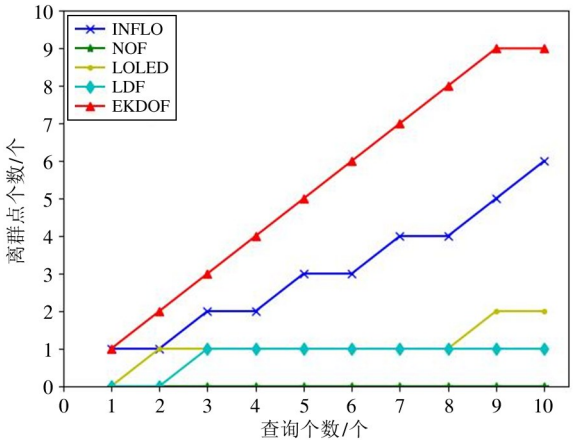
图 8 展示了 4 个真实数据集下,EKDOF 算法和 4 种对比算法离群点发现曲线的实验结果。



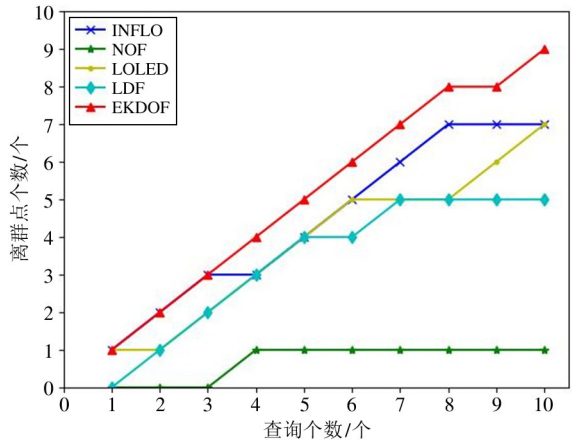
(a) PenDigits 离群点发现曲线



(b) Ecoli 离群点发现曲线



(c) WBC 离群点发现曲线



(d) wine 离群点发现曲线

图 8　不同算法在真实数据集上的离群点发现曲线对比

观察图 8(a) ~ (d),当用户的查询数量逐渐增加时,从各算法对应的离群点发现曲线的变化趋势可以得出,在 WBC 数据集和 wine 数据集上,EKDOF 算法离群点发现曲线的线性上升的速度与其他算法相比优势更加明显,突出了 EKDOF 算法在动态变化的查询条件下更能满足用户的期望。在 Ecoli 数据集上,除 NOF 算法整体表现较差外,本文算法和另外几种算法具有类似的性能,但整体上仍然优于其余对比算法。在 PenDigits 数据集上,本文所提出的 EKDOF 算法和 INFLO 算法的离群点发现曲线基本重合,但整体上本文算法离群点发现曲线的斜率更接近于 1,最终本文算法比 INFLO 算法多检测出 1 个离群点。LOLED 算法和 LDF 算法在 Ecoli 数据集上的表现都较为良好,但在其他 3 个数据集上的表现却差强人意。综合上述的分析,EKDOF 算法的离群点发现曲线在每个真实数据集上的表现都是最优的,从而验证了本文所提出的基于期望核密度离群因子的离群点检测算法具有良好的检测精度和稳定性。

4 结 论

本文分析了近年来较新颖的基于密度的离群点检测算法和相关思想,针对基于密度的方法存在的问题进行了深入研究,提出了一种基于期望核密度离群因子的离群点检测算法。使用 k 近邻和反向 k 近邻扩展邻域空间代替传统的 k 邻域空间,充分考虑数据点邻域范围内的数据分布;将传统核密度估计的方法与多元高斯函数相结合,引入自适应核带宽的思想,避免了人为设定核带宽,更好地适应不同数据集的数据分布。本文提出了期望距离,并定义了期望核密度离群因子刻画数据对象离群程度,从而检测出离群点。对所提算法的正确性和复杂性进行分析,在人工数据集和真实数据集上的实验证明,EKDOF 算法具有良好的检测精度和稳定性,能够在各种分布的数据集上表现出优越的性能。

参考文献

[1] BHATTACHARYA G, GHOSH K, CHOWDHURY A S. Outlier detection using neighborhood rank difference[J]. Pattern Recognition Letters, 2015, 60-61:24-31.

[2] 琚安康, 郭渊博, 李涛, 等. 基于网络通信异常识别的多步攻击检测方法[J]. 通信学报, 2019,40(7):57-66.

[3] 任家东, 刘新倩, 王倩, 等. 基于 KNN 离群点检测和随机森林的多层入侵检测方法[J]. 计算机研究与发展, 2019,56(3):566-575.

[4] REZAPOUR M. Anomaly detection using unsupervised methods. Credit card fraud case study[J]. International Journal of Advanced Computer Science and Applications, 2019,10(11):1-8.

[5] HUANG D, MU D, YANG L, et al. CoDetect:financial fraud detection with anomaly feature detection[J]. IEEE Access, 2018,6:19161-19174.

[6] STOJANOVIC N, DINIC M, STOJANOVIC L. A data-driven approach for multivariate contextualized anomaly detection: industry use[C]// 2017 IEEE International Conference on Big Data (Big Data). Boston, USA: IEEE, 2017:1560-1569.

[7] YOU L, PENG Q, XIONG Z, et al. Integrating aspect analysis and local outlier factor for intelligent review spam detection[J]. Future Generation Computer Systems, 2020,102:163-172.

[8] KIRLIDOG M, ASUK C. A fraud detection approach with data mining in health insurance[J]. Procedia Social and Behavioral Sciences, 2012,62:989-994.

[9] WU D F. A regression sequences based method for high dimensional outlier detection[J]. Journal of Discrete Mathematical Sciences and Cryptography, 2017,20(4):931-943.

[10] ATKINSON A, BARNETT V, LEWIS T. Outliers in statistical data[J]. Journal of the Royal Statal Society Series A (Stats in Society), 1995,158(3):630-630.

[11] YUAN Z, ZHAN X Y, FEND S. Hybrid data driven outlier detection based on neighborhood information entropy and its developmental measures[J]. Expert Systems with Applications, 2018, 112(12): 243-257.

[12] DASHDONDOV K, KIM M H. Mahalanobis distance based multivariate outlier detection to improve performance of hypertension prediction[J]. Neural Processing Letters, 2023,55:265-277.

[13] TANG X Q, ZHU P. Hierarchical clustering problems and analysis of fuzzy proximity relation on granular space[J]. IEEE Transactions on Fuzzy Systems, 2013, 21(5):814-824.

[14] 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):37-45.

[15] QIN X, CAO L, RUNDENSTEINER E A, et al. Scalable kernel density estimation-based local outlier detection over large data streams[C]// The 22nd International Conference on Extending Database Technology. Lisbon, Portugal: EDBT, 2019,3:421-432.

[16] LI K S, GAO X, FU S Y, et al. Robust outlier detection based on the changing rate of directed density ratio[J]. Expert Systems with Application, 2022,207:1-13.

[17] RAMASWAMY S, RASTOGI R, SHIM K. Efficient algorithms for mining outliers from large data sets[J]. SIGMOD Record, 2000,29(2):427-438.

[18] 朱庆生, 唐汇, 冯骥. 一种基于自然最近邻的离群点检测算法[J]. 计算机科学, 2014,41(3):276-278,305.

[19] 张忠平, 房春珍. 基于累积全熵的子空间聚类离群点检测算法[J]. 计算机集成制造系统, 2015,21(8):2249-2256.

[20] 张忠平, 邱敬仰, 刘丛, 等. 基于聚类离群因子和相互密度的离群点检测算法[J]. 计算机集成制造系统, 2019,25(9):2314-2323.

[21] BREUNIG M M, KRIEGEL H P, NG R T, et al. LOF: identifying density-based local outliers[J]. SIGMOD Record, 2000,29(2):93-104.

[22] HUANG J, ZHU Q, YANG L, et al. A non-parameter outlier detection algorithm based on natural neighbor[J]. Knowledge-Based Systems, 2016,92:71-77.
[23] LI K, GAO X, JIA X, et al. Detection of local and clustered outliers based on the density-distance decision graph [J]. Engineering Applications of Artificial Intelligence, 2022,110:104719-104734.
[24] WAHID A, RAO A C S, DEB K. A relative kernel-density based outlier detection algorithm[C]//The 12th International Conference on Software, Knowledge, Information Management and Applications. Phnom Penh, Cambodia : SKIMA, 2018:1-7.
[25] WAHID A, ANNAVARAPU C S R. NaNOD: a natural neighbour-based outlier detection algorithm[J]. Neural Computing and Applications, 2021,33(6):2107-2123.
[26] ZHU Q, FENG J, HUANG J. Natural neighbor: a self-adaptive neighborhoodmethod without parameter K[J]. Pattern Recognition Letters, 2016,80:30-36.
[27] PAVLIDOU M, ZIOUTAS G. Kernel density outlier detector[J]. Topics in Nonparametric Statistics, 2014,74: 241-250.
[28] LATECKI L J, LAZAREVIC A, POKRAJAC D. Outlier detection with kernel density functions[C]//The 5th International Conference on Machine Learning and Data Mining in Pattern Recognition. Leipzig, Germany: MLPRIS, 2007:61-75.
[29] WANG C, LIU Z, GAO H, et al. VOS: a new outlier detection model using virtual graph [J]. Knowledge-Based Systems, 2019,185:1-12.
[30] FRANK A, ASUNCION A. UCI machine learning repository[EB/OL]. [2023-06-20]. https://archive.ics.uci.edu/index.php.

Outlier detection algorithm based on expected kernel density outlier factor

ZHANG Zhongping[*,**], SUN Guangxu[*], YAO Chunchen[*], LIU Shuo[*], QI Wenxu[***]
([*]School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004)
([**]Key Laboratory for Computer Virtual Technology and System Integration of Hebei Province, Yanshan University, Qinhuangdao 066004)
([***]School of Information Systems Engineering, Information Engineering University, Zhengzhou 450001)

Abstract

For the problem that density-based outlier detection method has low detection accuracy on different distributed data sets, an outlier detection algorithm based on expected kernel density outlier factor is proposed. Firstly, the k-nearest neighbor and reverse k-nearest neighbor extended neighborhood space are introduced instead of the traditional k-neighborhood range, and the neighborhood information of data objects is considered more comprehensively. Then, the multivariate Gaussian function is introduced on the basis of the traditional kernel density estimation (KDE) method to estimate the density of data objects in the extended neighborhood space, and the idea of adaptive kernel bandwidth is introduced to better adapt to the data distribution of different datasets. In addition, the concept of expected distance is proposed to further distinguish between local outliers and normal points located in low-density regions. Finally, the expected kernel density outlier factor characterizes the degree of outlier of the data object. The proposed algorithm is experimentally verified on artificial datasets and real datasets, and compared with some traditional algorithms to prove the effectiveness of the proposed algorithm.

Key words: data minning, outlier, kernel density estimation (KDE), expected distance, expected kernel density outlier factor