ERDOF：基于相对熵权密度离群因子的离群点检测算法

张忠平[1,2,3]，刘伟雄[1]，张玉停[1]，邓禹[1]，魏棉鑫[1]

（1. 燕山大学信息科学与工程学院，河北 秦皇岛 066004；
2. 河北省计算机虚拟技术与系统集成重点实验室，河北 秦皇岛 066004；
3. 河北省软件工程重点实验室，河北 秦皇岛 066004）

摘 要：针对现有离群点检测算法在复杂数据分布和高维度数据集上精度低的问题，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。首先引入熵权距离取代欧氏距离以提高离群点检测精度。然后结合自然邻居的概念对数据对象进行高斯核密度估计。同时提出相对距离来刻画数据对象偏离邻域的程度，提高所提算法在低密度区域检测离群点的能力。最后提出相对熵权密度离群因子来刻画数据对象的离群程度。在人工数据集和真实数据集下进行的实验表明，所提算法能有效适应各种数据分布和高维数据的离群点检测。

关键词：数据挖掘；离群点检测；信息熵；核密度估计

中图分类号：TP311

文献标识码：A

DOI: 10.11959/j.issn.1000−436x.2021152

ERDOF: outlier detection algorithm based on entropy weight distance and relative density outlier factor

ZHANG Zhongping[1,2,3], LIU Weixiong[1], ZHANG Yuting[1], DENG Yu[1], WEI Mianxin[1]

1. College of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China
2. The Key Laboratory for Computer Virtual Technology and System Integration of Hebei Province, Qinhuangdao 066004, China
3. The Key Laboratory of Software Engineering of Hebei Province, Qinhuangdao 066004, China

Abstract: An outlier detection algorithm based on entropy weight distance and relative density outlier factor was proposed to solve the problem of low accuracy in complex data distribution and high dimensional data sets. Firstly, entropy weight distance was introduced instead of euclidean distance to improve the detection accuracy of outliers. Then, the Gaussian kernel density estimation was carried out for the data object based on the concept of natural neighbor. At the same time, relative distance was proposed to describe the degree of the data object deviating from the neighborhood and improve the ability of the algorithm to detect outliers in the low-density region. Finally, the entropy weight distance and relative density outlier factor were proposed to describe the degree of outliers. Experiments with artificial data sets and real data sets show that the proposed algorithm can effectively adapt to various data distributions and outlier detection of high-dimensional data.

Keywords: data mining, outlier detection, information entropy, kernel density estimation

1 引言

离群点是数据集中偏离大部分数据对象的数据，它们的表现和大多数数据对象有着明显的差异。离群点并不等同于错误数据，反而可能蕴含着极其重要的信息。离群点检测就是从海量数据中发现异常数据对象，是数据挖掘中的热门研究方向。目前，离群点检测广泛运用于工业无线传感器网络[1]、医疗处理[2]、欺诈检测[3]、垃圾邮件检测[4]、入侵检测[5-6]等领域，且有很多种分类，包括基于统计的[7-8]、基于距离的[9-10]、基于聚类的[11-13]和基于密度的[14-19]等方法。

基于统计的检测方法是根据现有的数据集结

收稿日期：2021−03−23；**修回日期**：2021−06−30

基金项目：河北省创新能力提升计划基金资助项目（No.20557640D）

Foundation Item: Hebei Province Innovation Capability Improvement Plan Project (No.20557640D)

合统计学的方法生成模型，这类方法根据数据对象在模型中是否处于低概率区域来判断离群点。这类检测方法需要充分的数据先验知识，并且对于高维数据集的检测效果较差。

基于距离的检测方法主要通过计算数据集中每个数据对象和邻居点的距离来检测离群点，其中 k 最近邻（KNN, k nearest neighbor）算法[9]是基于距离的方法中较常用的算法，基本原理是通过计算数据对象与其 k 个最近邻居的平均距离，通常离群点会远离正常点，因此平均距离越大，越有可能是离群点。由于 KNN 算法通过全局距离来检测离群点，因此无法检测出局部离群点。

基于聚类的检测方法是一种无监督的学习方法，用作聚类的数据集不需要类标签，基本思想是将数据集通过聚类算法得到簇，并将数据对象与簇进行比较，通常离群点不属于任何密集簇。常用的算法有 DBSCAN[12]和 CHAMELEON[13]。然而此类方法需要根据不同的应用场景和数据本身特征采用不同的聚类方法，因此检测的有效性依赖于聚类方法的选用。

在基于密度的检测方法中，如果一个数据对象为离群点，那它的密度与其周围邻居的密度会有很大的差异。这类密度检测方法主要通过数据对象的局部密度和其邻域密度的差异来判断离群点。为了实现这个想法，研究人员提出了很多基于密度的离群点检测算法。其中局部离群因子（LOF, local outlier factor）算法[14]是最常用的一种离群点检测算法。LOF 代表一个数据对象的离群得分，用于表示该对象与其局部可达邻域之间的差异。该方法利用 2 个数据对象之间可达距离来估计该对象的密度，该对象的离群得分是根据数据对象相对于其邻域对象的相对密度得出来的。文献[14]表明，离群得分更高的数据对象更有可能是离群点。之后，研究人员提出了很多基于 LOF 的改进算法。Zhang 等[16]提出了一种基于局部距离的离群因子（LDOF, local distance-based outlier factor）算法来度量离群得分。文献[17]在 LDOF 算法的基础上，增加了熵权距离的定义，改进并提出了一种基于熵权距离和局部密度的离群点检测（ELDOF, local entropy weight distance-based outlier factor）算法。

近年来，基于核的方法在离群点检测和聚类等领域得到了广泛的应用，基于核的方法利用核函数及其参数建立算法模型。文献[18]提出了一种基于密度的离群点检测算法，该算法将核密度估计（KDE, kernel density estimation）纳入 LOF 框架中，并将一个数据对象的 KDE 标准化成 S 分数，与其邻域的 KDE 进行比较。

与本文算法最相似的算法是基于自然邻居的离群点检测（NaNOD, natural neighbour-based outlier detection）算法[19]，该算法是一种非监督的基于密度的离群点检测算法，使用自然邻居概念，自适应地获取一个自然值（NV, natural value）的参数，并使用加权核密度估计（WKDE, weighted kernel density estimation）方法来估计数据对象的密度。另外，该算法采用了 KNN 以及反向 k 最近邻（RNN, reverse k-nearest neighbor），使系统可以灵活地对不同数据模式进行建模，并采用高斯核函数来实现度量的平滑性。此外，该算法使用自适应核宽度概念来增强正常样本与离群样本之间的区分能力。

分析近年来较新颖的基于密度的离群点检测算法和相关算法思想可以发现，大多数基于密度的检测算法在一些低密度的区域内和在高维数据集上的检测效果会有所下降。因此，本文在已有算法的基础上，考虑到在低密度区域内局部离群点与内部点密度相近的问题，提出相对距离的概念，可以有效地将其区分开来，从而提高算法在低密度区域处理局部离群点的能力。此外，本文还考虑到高维度对于离群点检测的影响，并不是所有属性对离群点检测都是有作用的，因此本文引用了信息熵加权距离（EWdist, entropy-weighted distance）的概念，取代传统算法中的欧氏距离，为数据对象的不同属性分配不同的权重，给离群属性分配更高的权重，能放大离群点的离群程度，从而提高算法在一些高维数据集中的检测能力。针对上述问题，本文提出了一种基于相对熵权密度离群因子的离群点检测（ERDOF, outlier detection based on entropy weight distance and relative density outlier factor）算法来检测离群点。

2 相关工作

2.1 自然邻居

Zhu 等[20]提出了一个新的无参数邻居的概念，称为自然邻居。如果数据对象 x 把数据对象 y 看作自己的邻居，同时 y 也把 x 看作自己的邻居，那么就把 x 看作 y 的自然邻居。一般地，在稀疏区域内的数据对象拥有较少的邻居，而在稠密区域内的数

据对象则拥有更多的邻居。

更重要的是，自然邻居可以在不使用任何参数的情况下有效地计算邻域。其主要思路是不断扩展邻居的搜索范围，每次搜索时记录每个数据对象被看作其他对象的邻居的次数，直到不被其他数据对象看作邻居的对象个数不变。由于数据对象在数据集中的 KNN 以及 RNN 的搜索代价较高，因此本文在自然邻居搜索算法[20]中使用 Ball-Tree[21]。Ball-Tree 是一个轻量级的二叉树，在高维数据集上性能良好且查询效率高。自然邻居搜索算法[20]流程如算法 1 所示。

算法 1　自然邻居搜索算法

输入　初始数据集 D

输出　自然近邻值 NV

定义 r=1，flag=0。初始化 $\mathrm{NaN}(x)=0$，$\mathrm{KNN}(x)=\varnothing$，$\mathrm{RNN}(x)=\varnothing$；

1) 根据数据集 D 创建一个 Ball-Tree；
2) 循环
3) while flag==0 do
4) 循环
5) for each $x\in D$ do
6) 用 Ball-Tree 找到 x 的前 r 个邻居 y；
7) $\mathrm{NaN}(y)=\mathrm{NaN}(y)+1$；
8) $\mathrm{KNN}_r(x)=\mathrm{KNN}_{r-1}(x)\cup\{y\}$；
9) $\mathrm{RNN}_r(y)=\mathrm{RNN}_r(y)\cup\{x\}$；
10) end for
11) $\mathrm{num}=\mathrm{count}(\mathrm{NaN}(x)==0)$；
12) if num 没有变化 then
13) flag=1；
14) end if
15) r=r+1；
16) end while
17) $\mathrm{NV}=\max(\mathrm{NaN}(x))$
18) 输出自然近邻值 NV

在算法 1 中，r 表示邻域搜索范围，$\mathrm{NaN}(x)$ 表示对象 x 被其他对象看作邻居的次数，$\mathrm{KNN}_r(x)$ 表示对象 x 的 r 最近邻域，$\mathrm{RNN}_r(y)$ 表示对象 y 的反向 r 最近邻域。本文在算法 1 中应用了 Ball-Tree 来提高邻域的搜索效率，时间复杂度为 $O(n\log n)$，其中 n 表示数据集中对象的数量。

定义 1　自然邻域。$\mathrm{KNN}_k(x_i)$ 和 $\mathrm{RNN}_k(x_i)$ 合并生成对象 x_i 的扩展邻域空间称为自然邻域，定义为 $\mathrm{IS}(x_i)=\mathrm{KNN}_k(x_i)\cup\mathrm{RNN}_k(x_i)$。其中，邻域的 k 值不是人为设定的，而是通过算法 1 自适应得出的自然近邻值 NV。

2.2　熵权距离

本文在计算数据对象之间的距离的时候，使用熵权距离取代传统算法中的欧氏距离，下面，给出信息熵加权距离的定义和信息熵加权算法[17]的流程描述。

设数据集 $D=\{x_1,x_2,\cdots,x_n\}$，其中 n 为数据样本大小。设属性集 $A=\{A_1,A_2,\cdots,A_d\}$，其中 d 为数据样本维度数量。信息熵表示信息的不确定程度。熵值越大，信息的不确定程度越大，信息熵计算方法如式(1)所示。

$$H(A_i)=-\sum_{x_j\in S(A_i)}p(x_j)\log p(x_j) \tag{1}$$

其中，$S(A_i)$ 是属性 $A_i(i=1,2,\cdots,d)$ 所有可能的取值的集合。

定义 2　离群属性。在数据集 D 中，若属性 A_i 的信息熵大于或等于数据集的平均信息熵，则称之为离群属性。

$$H(A_i)\geqslant\frac{\sum_i H(A_i)}{d} \tag{2}$$

通过式(2)的判断，符合条件的属性 A_i 看作离群属性。根据条件判断公式分类属性后，各属性的权重 w_i 取值如式(3)所示，其中 $l>1$。

$$w_i=\begin{cases}l, & A_i\text{为离群属性}\\ 1, & A_i\text{为非离群属性}\end{cases} \tag{3}$$

在数据集中，并非所有属性对离群点检测都是有作用的，因此，在计算距离时，为离群属性提供更大的权值能更好地突显离群点的离群程度，从而提高离群点和内部点的区分能力，其中参数 l 在本文中默认设定为 1.5。

定义 3　熵权距离。熵权距离是具有信息熵加权的欧氏距离。对象 o 与对象 p 的熵权距离定义如式(4)所示。

$$\mathrm{EWdist}(o,p)=\sqrt{\sum_{i=1}^{d}w_i(f_{A_i}(o)-(f_{A_i}(p))^2} \tag{4}$$

其中，$f_{A_i}(o)$ 为对象 o 在属性 A_i 上的取值，$w_i(i=1,2,\cdots,d)$ 为属性 A_i 相应的权值。

2.3　高斯核密度估计

设数据集 $D=\{x_1,x_2,\cdots,x_n\}$，其中 n 为数据样本大小。为了计算数据对象不同于其自然邻域的偏差程度，首先进行局部密度的估计。由于不确定数据

集的分布情况，本文选用无参数的高斯核密度估计（GKDE, Gaussian kernel density estimation）方法[22]来估计数据对象的密度，该方法使用自适应核宽度概念来提高离群点和内部点之间的区分能力和平滑正常样本点（内部点）之间的差异。

使用 GKDE 在随机样本 $x_1,x_2,\cdots,x_n(x_i\in D^d)$ 上进行局部密度估计，计算方法如式(5)所示。

$$\rho(x_i)=\sum_{j=1}^{n}\frac{1}{{h_j}^d}K(\frac{x_i-x_j}{h_j}) \tag{5}$$

其中，$K(\cdot)$ 表示核函数；h_j 表示对应数据对象 j 自适应得出的核带宽，用于控制密度估计的平滑度。本文选用零均值、单位标准差的多变量 d 维的高斯核函数，计算方法如式(6)所示。

$$K\left(\frac{x_i-x_j}{h_j}\right)=\frac{1}{(2\pi)^{\frac{d}{2}}}\exp\left(\frac{-\left\|x_i-x_j\right\|^2}{2(h_j)^2}\right) \tag{6}$$

其中，$\left\|x_i-x_j\right\|^2$ 表示从 x_i 到 x_j 的熵权距离。

本文选用的 GKDE 方法[22]只通过数据对象的邻域去估计对象 x_i 的局部密度，而不是通过整个数据集去估计。因为如果通过整个数据集去估计，有可能无法检测到局部的离群点。

局部离群点如图 1 所示。从图 1 可以看出，C_2 集合的点整体间距、密度和分散情况较一致，可以认为是同一个簇；虽然相较于 C_2 集合的点，C_1 集合的点较分散，但不难看出，C_1 集合的点也属于同一个簇。o_1、o_2 相对孤立，可以视作离群点或异常点。如果通过整个数据集去估计对象的密度，全局离群点 o_1 可以较容易地分离出来，但局部离群点 o_2 的密度与 C_1 簇中的点的密度相近，有可能无法检测到局部离群点 o_2，导致算法的准确率下降。

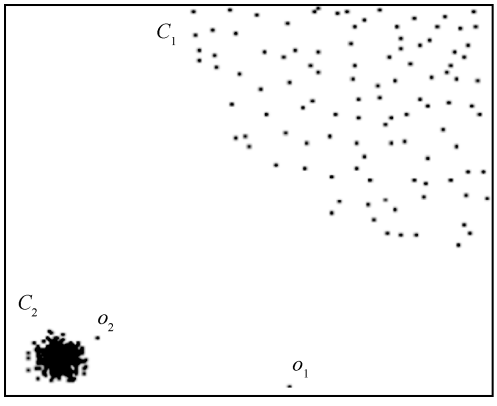


图 1 局部离群点

在基于全局的离群点检测算法中，KNN 算法[9]是较常用的算法，为了验证 KNN 算法在此类数据分布中的不足，本文使用 KNN 算法在图 1 数据集上进行离群得分的计算，结果如表 1 所示。KNN 算法中离群得分越大的点越有可能是离群点，而局部离群点 o_2 的离群得分比 C_1 簇的最大离群得分低，因此 KNN 算法无法检测出局部离群点 o_2。

表 1 KNN 算法在图 1 上的离群得分

数据集	离群得分
C_1 簇的最大值	0.547 4
C_2 簇的最大值	0.041 2
o_1	1.835 2
o_2	0.220 9

此外，通过整个数据集进行密度估计的计算成本较高（$O\left(n^2\right)$）。

为了更好地估计数据对象在邻居中的密度，本文使用数据对象 x_i 的 k 最近邻（$\rho_{\mathrm{IS}}(x_i)=\sum_{x\in \mathrm{IS}(x_i)}\frac{1}{{h_x}^d}K\left(\frac{x_i-x}{h_x}\right)$）和反向 k 最近邻（$\mathrm{RNN}_k$）的并集作为数据对象的邻域。其中数据对象 x_i 的 RNN_k 表示把 x_i 看作自己的 k 最近邻的集合，实验表明 RNN_k 可以更好地提供局部分布信息，将其用于检测离群点有较好的效果[23]。

因此，式(5)对于对象 x_i 的密度估计的计算如式(7)所示。

$$\rho_{\mathrm{IS}}(x_i)=\sum_{x\in \mathrm{IS}(x_i)}\frac{1}{{h_x}^d}K\left(\frac{x_i-x}{h_x}\right) \tag{7}$$

综合式(6)和式(7)，对象 x_i 的密度估计的计算如式(8)所示。

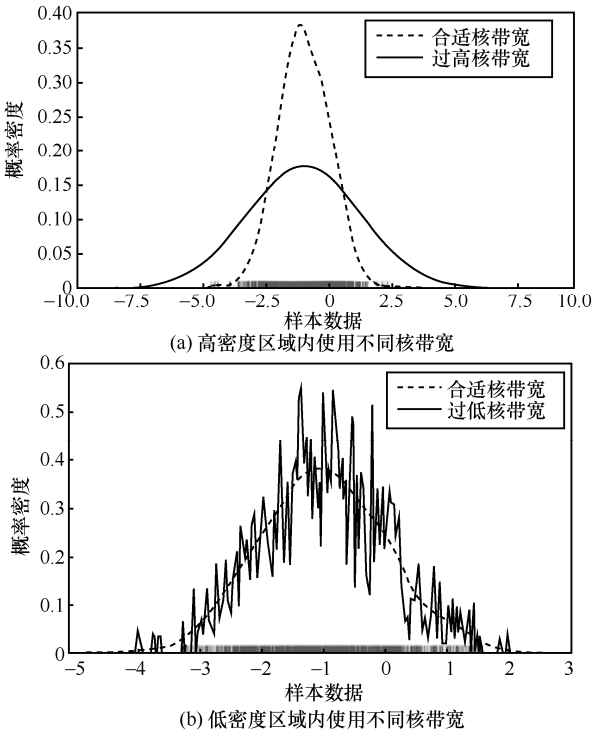
$$\rho_{\mathrm{IS}}(x_i)=\sum_{x\in \mathrm{IS}(x_i)}\frac{1}{{h_x}^d(2\pi)^{\frac{d}{2}}}\exp\left(\frac{-\left\|x_i-x\right\|^2}{2(h_x)^2}\right) \tag{8}$$

2.4 自适应核带宽

本文选用了高斯核密度估计方法[22]对数据对象进行密度估计，在传统的核密度估计方法里，核带宽都是预先设置好且固定的。但在使用的过程中会发现，在一些高密度区域里，使用过高的核带宽会使密度估计结果过于平滑，影响实验结果；此外，在一些低密度区域里，使用过低的核带宽会导致产生噪声估计。

图 2(a)表示在高密度区域设置不同核带宽对数据点密度估计的影响。当设置合适的核带宽时，各个数据点的密度为虚线所对应的曲线，能较准确地表示数据点的密度分布。当设置过高的核带宽时，各个数据点的密度为实线所对应的曲线，过高的核带宽会导致高密度区域的密度分布过于平滑，掩盖了数据大部分的基础结构，从而影响实验结果。

图 2(b)表示在低密度区域设置不同核带宽对数据点密度估计的影响。当设置合适的核带宽时，各个数据点的密度为虚线所对应的曲线，能较准确地表示数据点的密度分布。当设置过低的核带宽时，各个数据点的密度为实线所对应的曲线，过低的核带宽会导致低密度区域数据点的密度分布波动较剧烈，会产生大量噪声估计，从而影响实验结果。



(a) 高密度区域内使用不同核带宽

(b) 低密度区域内使用不同核带宽

图 2　核带宽对密度估计影响

因此本文需要为数据对象设置一个相对较优的核带宽，这一般取决于数据对象在数据集中所处的特定位置。为了获取这样的核带宽，本文引入自适应核带宽的概念[22]，目的是提高离群点和内部点的区分能力和平滑正常样本点（内部点）之间的差异。对于数据对象 x_i，取其 KNN_k 邻域内平均距离，记为 $d_{\mathrm{avg}}(x_i)$，如式(9)所示。

$$d_{\mathrm{avg}}\left(x_i\right)=\frac{1}{k}\sum_{j\in \mathrm{KNN}_k\left(x_i\right)} d\left(x_i,x_j\right) \tag{9}$$

此外，取数据集 $\left\{d_{\mathrm{avg}}\left(x_i\right)\mid i=1,2,3,\cdots,n\right\}$ 中的最大值和最小值，分别记为 $d_{\max}$ 和 $d_{\min}$。自适应核带宽 h_i 的计算方法如式(10)所示。

$$h_i=\theta\left[d_{\max}+d_{\min}+\delta-d_{\mathrm{avg}}\left(x_i\right)\right] \tag{10}$$

其中，$\theta(\theta>0)$ 是密度估计中用于控制平滑度的参数；δ 是一个非常小的正数，是为了防止核带宽取值为 0。

3　ERDOF 算法

3.1　相对距离

本文提出了一种相对距离的概念，记为 σ_i。考虑到在低密度区域内局部离群点与内部点的密度相近的问题，因此在计算数据对象的离群程度时，除了考虑数据对象的密度外，增加对数据对象相对距离的计算，进一步刻画数据对象的离群程度，能有效地把局部离群点和内部点、边界点区分开，从而提高算法在低密度区域处理局部离群点的能力。

定义 4　相对距离。相对距离是指数据对象到相对于密度比它大的数据对象的距离中的最小值。

$$\sigma_i=\min_{j:\rho_j>\rho_i}\left(d_{ij}\right) \tag{11}$$

对于密度最大的数据对象，本文通常给它取一个很小的正数作为该数据对象的相对距离。通常情况下，内部点处于一个密度较相近的区域内，相对距离的取值较小，而离群点常处于稀疏区域，密度比其大的点通常处于相对较远的位置，所以离群点的相对距离的值会比较大。通过该方法得出的相对距离，能够有效地提高 ERDOF 算法区分内部点和离群点的能力。

在完成对所有数据对象的密度估计以及相对距离的计算后，为了更好地刻画数据对象的离群程度，本文提出了相对熵权密度离群因子的概念，进一步刻画数据对象的离群程度，进而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 ERDOF 来检测数据集中的离群点。

3.2　相对熵权密度离群因子

定义 5　相对熵权密度离群因子。相对熵权密度离群因子是由相对距离和核密度的比值所构成的，计算方法如式(12)所示。

$$\mathrm{ERDOF}_k\left(x_i\right)=\frac{\sigma_i}{\rho_i} \tag{12}$$

该离群因子由相对距离和核密度的比值构成。从密度上看，本文首先通过计算自然邻居自适应获取 k 值，然后对数据集上的属性进行信息熵的计算，划分离群属性和非离群属性，为离群属性在计算距离时提供更大的权值。运用高斯核密度估计[22]结合自然邻居和熵权距离的概念计算每个数据对象的密度，为了在不同数据分布都能得到较优的密度估计，本文引用了自适应核带宽的概念自适应获取相对较优的核带宽。其中密度相对较高的对象通常处于簇的内部，而密度相对较低的对象则有可能为离群点。同时考虑到在一些低密度区域和边界上的数据对象难以仅凭密度的大小进行判断。因此，本文提出了相对距离的概念，在低密度区域的簇中的内部对象和边界上对象的相对距离会相对较小，而离群点的相对距离通常相对较大，可以有效地提高内部点和离群点的区分能力。本文通过距离和密度的比值的形式构成相对熵权密度离群因子，通常情况下相对熵权密度离群因子较大的点更有可能为离群点，因此算法选取排序后的离群因子中前 o 个点作为离群点输出，在算法实际应用中，o 的取值是根据实际数据集的情况人为设定的。在本文实验中，为了验证算法在各数据集上的有效性，o 的取值根据数据集已有的离群点标签个数设定。

算法 2 ERDOF 离群点检测算法

输入 初始数据集 D 和 Ball-Tree

输出 数据集 D 中前 o 个离群点

1) 用式(1)、式(2)和式(4)算出数据集 D 中各对象之间熵权距离；

2) 用算法 1 自适应获取 $k=\mathrm{NV}$；

3) 循环

4) for each $x \in D$ do

5) 用 Ball-Tree 计算 $\mathrm{KNN}_k(x)$；

6) 用 Ball-Tree 计算 $\mathrm{RNN}_k(x)$；

7) 创建局部邻域空间 $\mathrm{IS}_k(x)$（合并 $\mathrm{KNN}_k(x)$ 和 $\mathrm{RNN}_k(x)$）；

8) 用式(10)计算 x 的核带宽；

9) 用式(8)计算 x 的局部密度；

10) 用式(11)计算 x 的相对距离；

11) 用式(12)计算 x 的 $\mathrm{ERDOF}_k(x)$；

12) end for

13) 降序排列 $\mathrm{ERDOF}_k(x)$；

14) 输出前 o 个点作为离群点；

3.3 ERDOF 算法正确性

算法 2 详细描述了所提算法。其中熵权距离是基于信息熵计算得出的，在多维数据集中，并不是每个属性对检测离群点都是有帮助的，通过给属性分配权重能有效地提高检测精度。生成的扩展局部邻域 $\mathrm{IS}_k(x)$ 是由 $\mathrm{KNN}_k(x)$ 和 $\mathrm{RNN}_k(x)$ 合并而成的，其中 $\mathrm{RNN}_k(x)$ 能有效地提高对于局部离群点的检测精度。基于 $\mathrm{IS}_k(x)$ 对每个数据对象进行密度估计，然后根据密度分别计算每个数据对象的相对距离。最后通过计算相对熵权密度离群因子，对其进行降序排列，输出前 o 个离群点。

3.4 ERDOF 算法复杂性

ERDOF 算法的时间复杂度主要由以下 2 个部分组成：1) 为得到自适应的 NV 和自然邻域而构建的 Ball-Tree，时间复杂度为 $O(n\log n)$，其中 n 为数据集的数据对象的个数；2) 计算数据对象的相对熵权密度离群因子 $\mathrm{ERDOF}_k(x)$，时间复杂度为 $O(nk)$，其中 k 为 NV，因此 ERDOF 算法总的时间复杂度为 $O(n\log n)$。

4 实验与分析

为验证本文所提算法 ERDOF 在各种复杂数据分布上的性能，本节在人工数据集和真实数据集上进行实验验证。在实验中，将本文算法与常用的 6 种离群点检测算法（NaNOD[19]、IForest[24]、LDF[25]、RDOS[26]、NOF[27]、COPOD[28]）进行对比实验。实验环境如表 2 所示。

表 2 实验环境

软硬件环境	参数
CPU	2.60 Hz Inter i7-4720HQ
硬盘	512.0 GB
内存	16.0 GB
开发环境	PyCharm
编译环境	Python 3.8
可视化工具	PyCharm

4.1 算法有效性检测指标

在离群点检测实验中，大多数的数据集是高度不平衡的，即正常数据大量存在，而异常数据则十分稀有，这使准确率可能不适合作为离群点检测的性能指标。因此，本文使用接收器工作特性（ROC, receiver operating characteristics）曲线下方的面积

（AUC, area under curve）作为实验结果的评价指标，AUC 在离群点检测领域是最常见且有效的评价指标。ROC 曲线是真阳性率随假阳性率变化的曲线，其中真阳性率和假阳性率的定义分别如式(13)和式(14)所示。

$$\mathrm{TPR}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{13}$$

$$\mathrm{FPR}=\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} \tag{14}$$

其中，TP 表示预测为离群点且实际上也是离群点的个数；FP 表示预测为离群点但实际上是正常点的个数；TN 表示预测为正常点且实际上也是正常点的个数；FN 表示预测为正常点但实际上是离群点的个数。

AUC 的取值范围为 0~1，一个随机算法会产生一条近似于对角线的曲线（AUC=0.5），AUC 的取值越大，意味着在该算法里离群点的离群得分有更大的概率排在正常点之前[29]，因此 AUC 的取值越大，离群点检测算法效果越好。

F1 分数（F1-Score）是统计学中用来衡量二分类模型精确度的一种指标，它同时兼顾了分类模型的准确率和查全率。F1 分数可以看作模型准确率和查全率的一种加权平均，它的取值为 0~1。

在参数的选择上，本文选用各算法在相应文献中的默认值进行后续实验。在本文算法中，自适应核带宽的平滑参数 θ 设置为 0.015，为了防止核带宽为 0，将极小正数 δ 设置为 10^{-4}；在信息熵权距离中，离群属性的权值 l 设置为 1.5。

为了验证参数 θ 和参数 l 取值的有效性，本文选取了人工数据集 D_2 和真实数据集 Ionosphere 进行实验。其中人工数据集是二维数据集，数据集特征较相似，故选用样本个数和离群点比例均适中的 D_2 数据集；真实数据集 Ionosphere 的离群点占比较大，因此该数据集能更好地检测出参数的变化对实验结果的影响。

图 3 是本文算法在人工数据集 D_2 和真实数据集 Ionosphere 上采用不同自适应核带宽的平滑参数 θ 的准确度的实验结果。从图 3 可以看出，在人工数据集上，参数 θ 在 0.01~0.06 保持高效稳定；在真实数据集上，当参数 θ 大于 0.03 时准确率大幅度下降，因此选取 0.015 作为参数 θ 在本文算法上的默认取值。

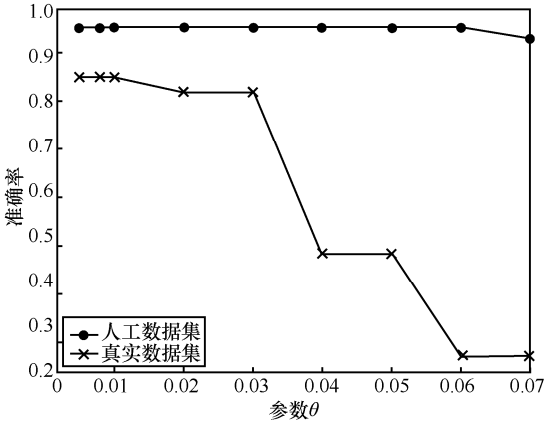


图 3　不同参数 θ 的准确度的实验结果

图 4 是本文算法在人工数据集 D_2 和真实数据集 Ionosphere 上采用不同的离群属性的权值参数 l 的准确度的实验结果。从图 4 可以看出，当离群属性的权重大于 1 时，在真实数据集上，准确率有了大幅提高，并且在 1.4~2.0 保持稳定，而信息熵加权距离主要适用于维度较高的数据集，人工数据集对参数 l 的设置并不敏感，因此选取 1.5 作为参数 l 在本文算法上的默认取值。

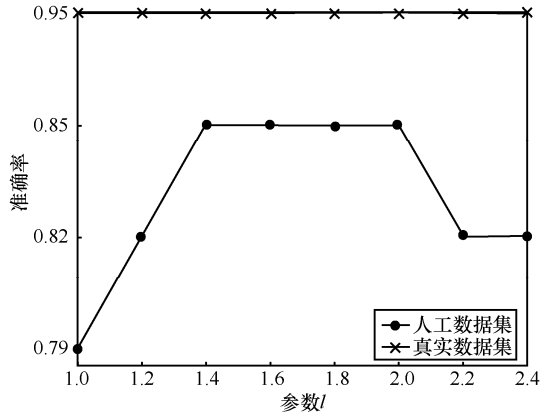


图 4　不同参数 l 的准确度的实验结果

4.2　人工数据集

为了验证本文算法在各种复杂数据分布下的性能，本文使用图 5 所示的 6 种二维人工数据集进行实验，其中离群点为“o”代表的点，人工数据集的数据特征如表 3 所示。

表 4 展示了在 6 种人工数据集上各算法的 AUC 得分结果，加黑字体表示每个数据集中表现最优的算法。在 D_4 中，本文 ERDOF 算法的 AUC 得分为 0.99，在所有对比算法中为最高得分。从表 4 中可以看出，本文 ERDOF 算法在所有数据集上的 AUC 得分基本都是最优的，与近年来较热门的算法相

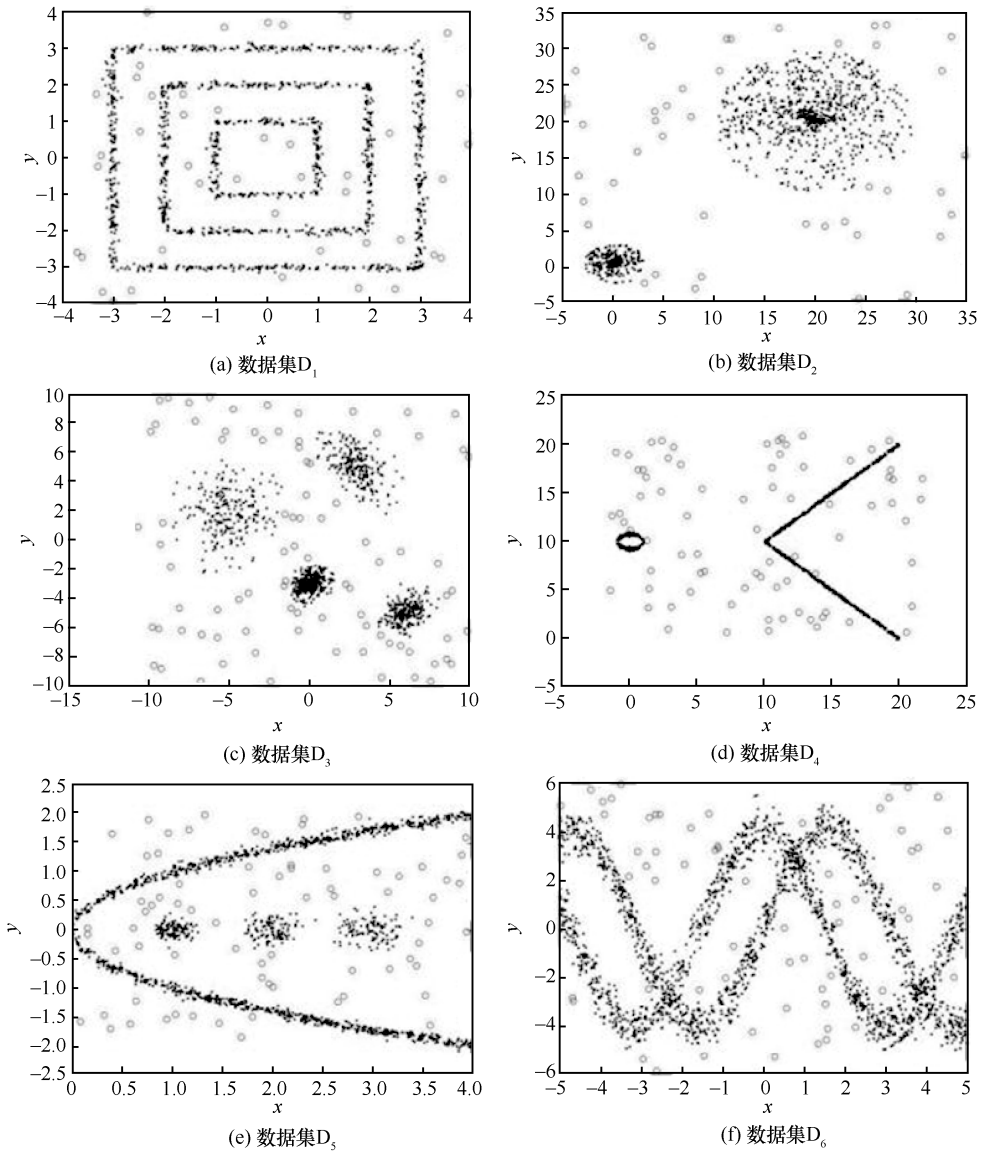


图 5　数据集 D_1~D_6 的数据分布

比，实验效果突出。虽然本文 ERDOF 算法并不在所有数据集上都表现最优，但整体性能远超其他对比算法。因此，该实验证明 ERDOF 算法可以适应各种复杂形状的数据分布且有较好的性能表现。

表 3　　人工数据集数据特征

数据集	样本个数/个	离群点个数/个	离群点比例
D_1	1 256	43	3.4%
D_2	1 043	43	4.1%
D_3	1 000	85	8.5%
D_4	876	77	8.7%
D_5	1 372	72	5.2%
D_6	2 042	64	3.1%

表 4　　各算法在人工数据集上的 AUC 得分

算法	D_1	D_2	D_3	D_4	D_5	D_6
ERDOF	**0.99**	**0.99**	**0.96**	**0.99**	**0.98**	**0.96**
NaNOD	**0.99**	**0.99**	0.95	0.97	0.96	0.89
NOF	0.97	0.67	0.85	0.74	0.92	0.92
IForest	0.93	**0.99**	0.89	0.90	0.68	0.82
LDF	0.94	0.95	0.93	0.97	0.95	0.81
RDOS	0.89	0.95	0.86	0.98	0.97	0.89
COPOD	0.67	0.86	0.92	0.16	0.74	0.35

4.3　真实数据集

本文采用的 6 种真实数据集均来自 UCI 数据集，数据集的维度为 4~166，离群点所占比例为

4.1%~35.8%，从维度和离群点占比上全面检测 ERDOF 算法的真实有效性，其中真实数据集的数据特征如表 5 所示。

表 5　　真实数据集数据特征

数据集	样本个数/个	属性个数/个	离群点个数/个	离群点比例
Iris	110	4	10	9.1%
Wdbc	390	30	33	8.4%
Wbc	223	9	10	4.5%
Ionosphere	351	34	126	35.8%
Rbds	372	105	38	10.2%
Musk	5 852	166	241	4.1%

图 6 是各算法在真实数据集上的离群点检测准确率。为了提高实验结果的可靠性，引入 F1 得分的概念对算法效果进行评测，结果如表 6 所示。从图 6 可以看出，ERDOF 算法在各个真实数据集上的准确率均不低于 0.8，ERDOF 算法与所有对比算法相比，在真实数据集上的准确率整体上达到最高。ERDOF 算法在真实数据集 Wbc 上的准确率和 F1 值均略低于 IForest 算法，但均高于与 NaNOD 算法和 NOF 算法。通过分析发现，真实数据集 Wbc 属性列密度分布都较均匀，使各个属性列的信息熵都较接近，进而导致熵权距离效果不佳，但由于本文使用相对距离去刻画密度分布均匀或低密度区域的密度分布，使 ERDOF 算法依旧保持较好的检测效果。其中在 Wdbc 和 Ionosphere 这 2 个高维数据集上，ERDOF 算法保持了较高的准确率，其余算法受维度灾难的影响，效果相对较差。ERDOF 算法在 Wdbc 数据集上的 F1 得分为 0.92，在所有对比算法中为最高得分。从表 6 中可以看出，ERDOF 算法在各个真实数据集上整体性能远超其他对比算法。实验验证了本文算法能全面准确地检测出离群点。

各算法在真实数据集上的运行时间如图 7 所示。从图 7 可以看出，在中低维度的数据集中各算法的运行时间基本持平，而在高维数据集中除了 COPOD 和 IForest 算法的运行时间保持平稳外，其余算法均大幅度增大。但综合准确率来看，除了 ERDOF 算法和 LDF 算法在高维数据集上保持了较高的准确率，其余算法的准确率均大幅降低。ERDOF 算法在时间效率上对比其余算法没有太大的优化，但在可接受的时间差范围内，为了提高离群点的检测精度，本文选用检测性能更好的熵权距离和相对距离，因此牺牲了一些时间效率，但准确率有较大提高。

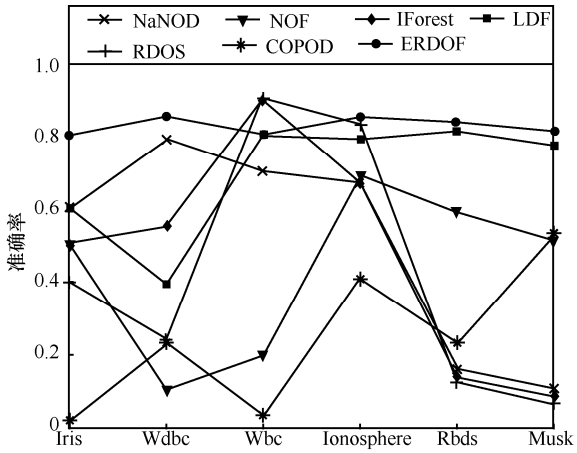


图 6　各算法在真实数据集上的离群点检测准确率

表 6　　各算法在真实数据集上的 F1 得分

算法	Iris	Wdbc	Wbc	Ionosphere	Rbds	Musk
ERDOF	**0.89**	**0.92**	0.90	**0.88**	**0.84**	**0.81**
NaNOD	0.78	0.88	0.84	0.75	0.16	0.11
NOF	0.73	0.52	0.58	0.76	0.59	0.51
IForest	0.73	0.75	**0.95**	0.83	0.81	0.77
LDF	0.78	0.67	0.90	0.83	0.81	0.77
RDOS	0.64	0.70	**0.95**	0.86	0.12	0.07
COPOD	0.84	0.80	0.89	0.10	0.54	0.49

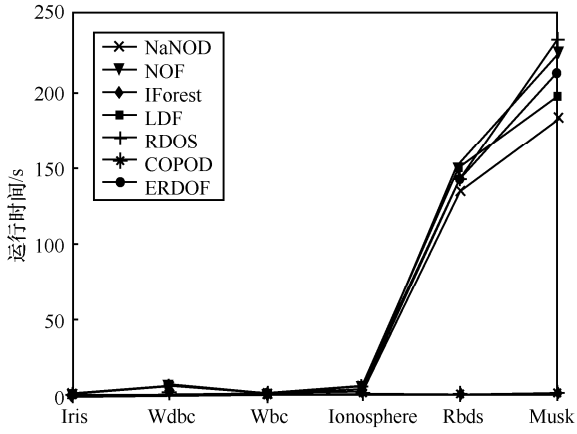


图 7　各算法在真实数据集上的运行时间

从图 6 和图 7 中可以看出，本文算法在维度 4~34 时，时间效率和准确率都保持较高的水平，而随着维度的增大，准确率依旧保持较高的水平，且时间效率的优势大幅度增加，因此本文算法考虑时间效率和准确率，在可接受的时间差范围内，本文算法可以应对

的维度的大致范围在 2~100。

5 结束语

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算法思想，针对基于密度的方法存在的问题，提出了相对熵权密度离群因子来刻画数据对象的离群程度，进而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，其中用熵权距离取代传统的欧氏距离，提高离群属性在计算距离中的权重。本文首先提出相对距离的概念，提高算法在低密度区域处理局部离群点的能力；然后对算法进行了正确性、复杂性的分析；最后在人工数据集和真实数据集上对 ERDOF 算法进行实验验证，通过对实验结果的比较分析，验证了 ERDOF 算法能有效且全面地检测离群点。

参考文献：

[1] RAMOTSOELA D, ABU-MAHFOUZ A, HANCKE G. A survey of anomaly detection in industrial wireless sensor networks with critical water system infrastructure as a case study[J]. Sensors, 2018, 18(8): 2491.

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

[3] ANDRYSIAK T. Sparse representation and overcomplete dictionary learning for anomaly detection in electrocardiograms[J]. Neural Computing and Applications, 2020, 32(5): 1269-1285.

[4] 杨加, 李笑难, 张扬, 等. 基于大数据分析的校园电子邮件异常行为检测技术研究[J]. 通信学报, 2018, 39(S1): 116-123.
YANG J, LI X N, ZHANG Y, et al. Abnormal behavior detection for campus email systems based on big data analysis[J]. Journal on Communications, 2018, 39(S1): 116-123.

[5] DENNING D E. An intrusion-detection model[J]. IEEE Transactions on Software Engineering, 1987, 13(2): 222-232.

[6] 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.
JU A K, GUO Y B, LI T, et al. Multi-step attack detection method based on network communication anomaly recognition[J]. Journal on Communications, 2019, 40(7): 57-66.

[7] ROUSSEEUW P J, LEROY A M. Robust regression and outlier detection[M]. New York: John Wiley & Sons, Inc., 1987.

[8] BARNETT V, LEWIS T, ABELES F. Outliers in statistical data[M]. 3rd ed. Hoboken: John Wiley & Sons, 1994.

[9] KNORR E M, NG R T, TUCAKOV V. Distance-based outliers: algorithms and applications[J]. The VLDB Journal, 2000, 8(3/4): 237-253.

[10] KNORR E M, NG R T. A unified approach for mining outliers: properties and computation[C]// Proceedings of Conference of the Centre for Advanced Studies on Collaborative Research. [S.n.:s.l.], 1997: 219-222.

[11] JAIN A K, MURTY M N, FLYNN P J. Data clustering[J]. ACM Computing Surveys, 1999, 31(3): 264-323.

[12] ESTER M, KRIEGEL, H, SANDER J, et al. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise[C]//International Conference on Knowledge Discovery & Data Mining. New York: ACM Press, 1996: 226-231.

[13] KARYPIS G, HAN E H, KUMAR V. Chameleon: hierarchical clustering using dynamic modeling[J]. Computer, 1999, 32(8): 68-75.

[14] 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. New York: ACM Press, 2000: 93-104.

[15] 杨晓晖, 刘晓明. 基于双向邻居修正的局部异常因子算法[J]. 通信学报, 2020, 41(8): 130-140.
YANG X H, LIU X M. Local outlier factor algorithm based on correction of bidirectional neighbor[J]. Journal on Communications, 2020, 41(8): 130-140.

[16] ZHANG K, HUTTER M, JIN H D. A new local distance-based outlier detection approach for scattered real-world data[C]//Advances in Knowledge Discovery and Data Mining. Berlin: Springer, 2009: 813-822.

[17] WANG L N, FENG C, REN Y J, et al. Local outlier detection based on information entropy weighting[J]. International Journal of Sensor Networks, 2019, 30(4): 207.

[18] SCHUBERT E, ZIMEK A, KRIEGEL H P. Generalized outlier detection with flexible kernel density estimates[C]//Proceedings of the 2014 SIAM International Conference on Data Mining. [S.n.:s.l.], 2014: 542-550.

[19] WAHID A, ANNAVARAPU C S R. NaNOD: a natural neighbour-based outlier detection algorithm[J]. Neural Computing and Applications, 2021, 33(6): 2107-2123.

[20] ZHU Q S, FENG J, HUANG J L. Natural neighbor: a self-adaptive neighborhood method without parameter K[J]. Pattern Recognition Letters, 2016, 80: 30-36.

[21] OMOHUNDRO S M. Five Balltree construction algorithms[R]. Technical Report, International Computer Science Institute, 1989.

[22] ZHANG L W, LIN J, KARIM R. Adaptive kernel density-based anomaly detection for nonlinear systems[J]. Knowledge-Based Systems, 2018, 139: 50-63.

[23] JIN W, TUNG A K H, HAN J W, et al. Ranking outliers using symmetric neighborhood relationship[C]//Advances in Knowledge Discovery and Data Mining. Berlin: Springer, 2006: 577-593.

[24] 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.

[25] LATECKI L J, LAZAREVIC A, POKRAJAC D. Outlier detection with kernel density functions[C]//Machine Learning and Data Mining in Pattern Recognition. Berlin: Springer, 2007: 61-75.

[26] TANG B, HE H B. A local density-based approach for outlier detection[J]. Neurocomputing, 2017, 241: 171-180.

[27] HUANG J L, ZHU Q S, YANG L J, et al. A non-parameter outlier detection algorithm based on Natural Neighbor[J]. Knowledge-Based Systems, 2016, 92: 71-77.

[28] LI Z, ZHAO Y, BOTTA N, et al. COPOD: copula-based outlier detec-

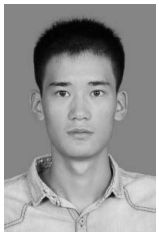
tion[C]//2020 IEEE International Conference on Data Mining. Piscataway: IEEE Press, 2020: 1118-1123.

[29] FLACH P A. Putting things in order: on the fundamental role of ranking in classification and probability estimation[C]//European Conference on Principles of Data Mining & Knowledge Discovery. Berlin: Springer, 2007: 2-3.

[作者简介]

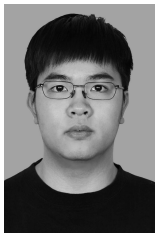
张忠平（1972− ），男，吉林松原人，博士，燕山大学教授，主要研究方向为大数据、数据挖掘、半结构化数据等。

刘伟雄（1997− ），男，广东广州人，燕山大学硕士生，主要研究方向为数据挖掘。



张玉停（1996− ），男，安徽阜阳人，燕山大学硕士生，主要研究方向为数据挖掘。

邓禹（1996− ），男，河北唐山人，燕山大学硕士生，主要研究方向为数据挖掘。



魏棉鑫（1997− ），男，广东汕头人，燕山大学硕士生，主要研究方向为数据挖掘。