ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 37,No. 5,Sep. 2022,pp. 1018－1025
DOI:10. 16337/j. 1004-9037. 2022. 05. 007

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

一种基于邻域近似精度的离群点检测方法

张玉婷[1]，冯　山[2]

(1. 四川师范大学计算机科学学院,成都 610068;2. 四川师范大学数学科学学院,成都 610068)

摘　要: 针对混合属性离群点检测问题,提出基于邻域近似精度的混合属性离群点检测方法。首先,定义异构邻域关系度量来表示混合数据之间的近邻性。然后,定义一种特定的邻域近似精度来构建邻域粒离群度。进而,定义基于邻域近似精度的离群因子及提出基于邻域近似精度的离群点检测(Nighborhood approximation accuracy-based outlier detection, NAAOD)。最后,用UCI数据集对NAAOD算法的有效性进行了验证。理论研究和实验结果均表明,NAAOD算法对混合属性离群点检测是有效的。

关键词: 离群点检测;领域精糙集;粒计算;邻域近似精度;混合属性

中图分类号: TP18　　**文献标志码:** A

An Outlier Detection Method Based on Neighborhood Approximate Accuracy

ZHANG Yuting[1], FENG Shan[2]

(1. School of Computer Science, Sichuan Normal University, Chengdu 610068, China;2. School of Mathematical Sciences, Sichuan Normal University, Chengdu 610068, China)

Abstract: Aiming at the problem of outlier detection of mixed attributes, this paper proposes a method for outlier detection of mixed attributes based on neighborhood approximate accuracy. First, a heterogeneous neighborhood relationship metric is defined to represent the proximity between mixed data. Then, a specific neighborhood approximation accuracy is defined to construct the neighborhood grain outliers. Further, a neighborhood approximation accuracy-based outlier factor is defined and a neighborhood approximation accuracy-based outlier detection (NAAOD) algorithm is proposed. Finally, the effectiveness of the NAAOD algorithm is evaluated using the UCI dataset. Theoretical research and experimental results show that the NAAOD algorithm is effective for detecting outliers with mixed attributes.

Key words: outlier detection; neighborhood rough sets; granular computing; neighborhood approximation accuracy; mixed attribute

引　　言

离群点检测多用于入侵检测、信用卡诈骗和医学诊断等领域[1-2],有分布式[3]、深度[4]、距离[5-6]、密度[7]和聚类[8]等方法。分布式检测要预设数据分布规律,不适合分布未知情形;深度法针对高维数据,效率低;距离和密度法由于利用欧式距离来设计,所以不是检测标称或混合属性离群点的最佳方法;聚

基金项目: 国家自然科学基金(61673285);四川省青年科技基金(2017JQ0046);四川省教育厅自然科学基金(15ZB0029)。
收稿日期: 2020-08-05;**修订日期:** 2021-05-06

类法开销大。

粒计算作为一个重要的研究方向,它主要分为2类:(1)以处理不确定性为主要目标,如以模糊处理和粗糙集为基础的计算模型;(2)以多粒度计算为目标,如商空间理论。其中,经典粗糙集理论模型[9]已经成功应用于标称特征选择和相关性分析等研究。近年来,针对距离和密度法不能有效处理包含标称属性数据的不足,提出了多种基于粗糙集的方法。例如,基于经典粗糙集,文献[10]采用粒计算的思想构建对象离群因子并做离群检测。文献[11]用粗糙边界定义对象异常度以做离群检测。文献[12]用粗糙集隶属度扩展了一种新方法。文献[13]提出了基于粗糙边界和距离的离群检测。文献[14]提出了基于经典近似精度的离群检测。但它们采用等价关系的方式建立数学模型,其检测模型均只适于处理标称属性数据集。

基于邻域粗糙集的特征选择[15-16]能有效处理混合属性数据集。例如,文献[17]提出了基于邻域模型的邻域离群检测。然而,其对参数选择非常敏感。文献[18-21]提出了面向混合属性数据的离群检测。但这些方法均未考虑邻域粗糙度量的离群点检测模型。邻域粗糙度量主要包括邻域近似精度和邻域粗糙度量等[16],它是度量混合属性数据不确定性的有效方法,可以用于混合属性数据集的离群点检测建模。

针对混合属性离群点检测问题,本文构造了一种基于邻域近似精度的离群点检测方法(Neighborhood approximation accuracy-based outlier detection, NAAOD)。该方法以优选异构邻域关系度量和统计邻域半径构建邻域信息系统(Neighborhood information system, NIS),以邻域近似精度离群因子表征对象离群度。对比实验表明,NAAOD算法能同时适于各种属性组合的离群检测。

1 预备知识

假设信息系统$I_S=(U,A,V,f)$,其中$U=\{x_1,x_2,\cdots,x_n\}$,属性集A非空;$V=\bigcup_{a\in A}V_a$,V_a为属性a的值域;$f:U\times A\to V$,$\forall a\in A$和$\forall x\in U$,$f(x,a)\in V_a$。当$A=C\cup D$时,信息系统变为决策系统,简记为$D_S=(U,C\cup D,V,f)$,其中$C=\{c_1,c_2,\cdots,c_m\}$且$D=\{d\}$。为讨论方便,设$B\subseteq C$且$|\cdot|$为集合的势。

定义1[15] 论域对象x_i的B邻域$\forall x_i\in U$和$\forall B\subseteq C$,$\delta_B(x_i)=\{x_j|x_j\in U,\Delta_B(x_i,x_j)\leqslant\varepsilon\}$是$x_i$的$B$邻域,$\varepsilon$是邻域半径,$\Delta_B(x_i,x_j)$是距离函数。对$\forall x_i,x_j,x_k\in U$,有:

(1) $\Delta_B(x_i,x_j)\geqslant 0\Leftrightarrow x_i=x_j,\Delta_B(x_i,x_j)=0$;

(2) $\Delta_B(x_i,x_j)=\Delta_B(x_j,x_i)$;

(3) $\Delta_B(x_i,x_k)\leqslant\Delta_B(x_i,x_j)+\Delta_B(x_j,x_k)$。

例如,关于条件属性子集C的混合欧式重叠度量(Heterogeneous euclidean overlap metric, HEOM)[22]为:$\mathrm{HEOM}(x,y)=\sqrt{\sum_{j=1}^{m}\omega_{\{c_j\}}\Delta^2_{\{c_j\}}(x,y)}$。其中

$$\Delta_{\{c_j\}}(x,y)=\begin{cases}1 & f(x,c_j)\text{或}f(y,c_j)\text{未知时}\\ 0 & c_j\text{为标称属性且}f(x,c_j)=f(y,c_j)\text{时}\\ 1 & c_j\text{为标称属性且}f(x,c_j)\neq f(y,c_j)\text{时}\\ \dfrac{\left|f(x,c_j)-f(y,c_j)\right|}{\max_{c_j}-\min_{c_j}} & c_j\text{是数值属性时}\end{cases}\tag{1}$$

式中$\omega_{\{c_j\}}$为c_j的权重。显然,式(1)可同时处理数值、标称属性及属性值未知情形。$\delta_B(x_i)$为关于属性

子集B的以x_i为中心的邻域粒,$\varepsilon=0$时退化为等价类。

$\forall x_i\in U,\boldsymbol{\delta}(x_i)\notin\varnothing$且$\bigcup\limits_{x\in U}\delta_B(x)=U$,称$\left\{\boldsymbol{\delta}_B(x_i)|x_i\in U\right\}$覆盖论域$U$。$U$上邻域关系$N_B$可写成关系矩阵$\boldsymbol{M}(N_B)=(r_{ij}^B)_{n\times n}$,$r_{ij}^B=\begin{cases}1 & \Delta_B(x_i,x_j)\leqslant\varepsilon\\ 0 & \text{其他}\end{cases}$。易知$N_B$是相似关系,邻域内样本彼此相似,$\boldsymbol{\delta}_B(x_i)$是0-1向量,即$\boldsymbol{\delta}_B(x_i)=(r_{i1}^B,r_{i2}^B,\cdots,r_{in}^B)$。

定义2[15] 论域子集的B邻域上、下近似$\forall X\subseteq U$,X的B-下近似为$\underline{N_B}X=\{x_i|\boldsymbol{\delta}_B(x_i)\subseteq X,x_i\in U\}$,而$B$-上近似为$\overline{N_B}X=\{x_i|\boldsymbol{\delta}_B(x_i)\cap X\neq\varnothing,x_i\in U\}$。

邻域粗糙集不确定性由边界域引起。因$\underline{N_B}X\subseteq X\subseteq\overline{N_B}X$,$X$边界域$BNX=\overline{N_B}X-\underline{N_B}X$,边界域大时系统精确性低。

定义3[16] 论域子集的B邻域近似精度$\forall X\subseteq U$,X的B-邻域近似精度即$\alpha_B(X)=\left|\underline{N_B}X\right|\Big/\left|\overline{N_B}X\right|$。

显然,$0\leqslant\alpha_B(X)\leqslant1$。$\alpha_B(X)=1$时$X$的$B$-边界域为空,$\alpha_B(X)=0\Leftrightarrow\underline{N_B}X=\varnothing$。易知,$X=\varnothing$时$\alpha_B(X)=1$。可用来反映论域集合中的知识的不确定性,可以被用于离群点检测模型的构建。

2 基于邻域近似精度的混合离群点检测

2.1 检测方法

可用最小-最大归一化[18]对数据集属性和维数差异作量纲统一。同HEOM,异构邻域关系度量(Heterogeneous neighborhood relation metric, HNRM)可定义如下:

定义4 对给定,$\forall x_i,x_j\in U,B=\{c_{j_1},c_{j_2},\cdots,c_{j_l}\}(1\leqslant l\leqslant m)\subseteq C$,$x_i$与$x_j$的混合或HNRM值$r_{ij}^B=\mathrm{HNRM}_{ij}^B=\min\limits_{h=1}^{l}r_{ij}^{c_{j_h}}$,其中

$$r_{ij}^{c_{j_h}}=\begin{cases}1 & c_{j_h}\text{为标称属性且}f(x,c_{j_h})=f(y,c_{j_h})\\ 0 & c_{j_h}\text{为标称属性且}f(x,c_{j_h})\neq f(y,c_{j_h})\\ 1 & \left|f(x,c_{j_h})-f(y,c_{j_h})\right|\leqslant\varepsilon\text{且}c_{j_h}\text{为数值属性}\\ 0 & \left|f(x,c_{j_h})-f(y,c_{j_h})\right|>\varepsilon\text{且}c_{j_h}\text{为数值属性}\end{cases}\tag{2}$$

易知,式(2)可同时度量不同属性组合的对象差异度,但ε一般由专家确定[17,19,21],其主观随意性会导致算法对参数敏感。为此,文献[18]提出了具有自适应特性的ε取值法,即

$$\varepsilon_{c_j}=\begin{cases}0 & c_j\text{为标称属性}\\ \dfrac{\mathrm{std}(c_j)}{\lambda} & c_j\text{为数值属性}\end{cases}\tag{3}$$

式中:$\mathrm{std}(c_j)$为属性标准差;λ为调整参数。显然,式(3)的邻域半径ε取值方式更为合理,它用于调整邻域半径。

性质1 条件属性子集并的邻域关系等价于其邻域关系的交,即对$\forall C_1,C_2\subseteq C$,$N_{C_1\cup C_2}=N_{C_1}\cap N_{C_2}$。

假设$C_1,C_2\subseteq C$分别为数值型和标称型条件属性子集。由性质1易知,对象x的邻域有如下等式成立:

(1) $\delta_{C_1}(x)=\bigcap\limits_{c_1\in C_1}\delta_{c_1}(x)$;

(2) $\delta_{C_2}(x)=\bigcap\limits_{c_2\in C_2}\delta_{c_2}(x)$;

(3) $\delta_{C_1\cup C_2}(x)=\delta_{C_1}(x)\cap\delta_{C_2}(x)$。

对$x\in U$和U上邻域关系,可得x的一组邻域粒。利用x的邻域粒差异或距离等信息可以进行离群检测。一般而言,为计算x的离群度,可先算邻域粒离群度,再确定离群因子。邻域粒离群度可用邻域粒的邻域近似精度度量。

定义 5 特定邻域近似精度 $\forall B\subseteq C$, $G=\{\boldsymbol{\delta}_B(x_i)|x_i\in U\}$, $E\subseteq C-B$, $\boldsymbol{\delta}_B(x_i)$关于$N_E$的邻域近似精度可定义为

$$\alpha_E\left[\boldsymbol{\delta}_B(x_i)\right]=\frac{\left|\underline{N_E}\boldsymbol{\delta}_B(x_i)\right|}{\left|\overline{N_E}\boldsymbol{\delta}_B(x_i)\right|} \tag{4}$$

邻域近似精度常用来度量决策类的论域近似度。一般来说,$\boldsymbol{\delta}_B(x_i)$对一组邻域关系近似精度很低时表明$\boldsymbol{\delta}_B(x_i)$行为异常或离群程度高。而经典近似精度离群因子[14]仅适于度量标称属性。对混合属性,可用邻域粒离群度(Neighborhood granule outlier degree, NGOD)和邻域近似精度离群因子(Neighborhood approximation accuracy-based outlier factor , NAAOF)有效度量。前者度量给定邻域粒的离群程度,后者度量给定对象的离群程度。

定义 6 令$E=C-B=\{c_{e_1},c_{e_2},\cdots,c_{e_k}\}(|C-B|\geqslant 2)$, $\boldsymbol{\delta}_B(x_i)\in G$关于$N_B$的NGDD可定义为

$$\mathrm{NGOD}\left(\boldsymbol{\delta}_B(x_i)\right)=1-\frac{\left|\boldsymbol{\delta}_B(x_i)\right|}{|U|}\cdot\frac{\alpha_E\left[\boldsymbol{\delta}_B(x_i)\right]+\sum\limits_{j=1}^{k}\alpha_{E-\{c_{e_j}\}}\left[\boldsymbol{\delta}_B(x_i)\right]}{k+1} \tag{5}$$

式中$\boldsymbol{\delta}_B(x_i)$对一组邻域关系的近似精度很低时$\boldsymbol{\delta}_B(x_i)$对邻域关系影响小,可认为$\boldsymbol{\delta}_B(x_i)$行为异常且离群度高。此时,NGOD($\boldsymbol{\delta}_B(x_i)$)大,可以度量$\boldsymbol{\delta}_B(x_i)$的异常性,进而度量$x_i$的离群性。基于NGOD的邻域近似精度离群因子,即论域对象x_i的NAAOF定义如下:

定义 7 $x_i\in U$的邻域近似精度离群因子可定义为

$$\mathrm{NAAOF}(x_i)=\frac{\sum\limits_{j=1}^{m}\left(\mathrm{NGOD}(\delta_{\{C_j\}}(x_i))\times W_{\{C_j\}}(x_i)\right)}{|C|} \tag{6}$$

权重函数$W_{\{C_j\}}(x_i)=1-\sqrt[3]{|\delta_{\{C_j\}}(x_i)|/|U|}$的映射取值范围为$W_{\{C_j\}}: U\to[0,1]$。$x_i\in U$和$C_j\in C$,$|\delta_{C_j}(x_i)|$小于相关的其他邻域粒子时,$x_i$可被视为$U$中少数对象,应被赋予更高的离群权重值。$W_{\{C_j\}}(x_i)$值越高,$x_i$的邻域近似精度离群因子NAAOF($x_i$)越大。对$|\delta_{\{C_j\}}(x_i)|/|U|$开立方根可提升邻域粒近似精度的影响力。因此,对给定离群阈值μ和$x\in U$, NAAOF(x)$>\mu$时x被当作邻域近似精度离群点。

2.2 检测算法步骤及复杂度

算法步骤如下:

输入:$I_S=(U,C,V,f)$, 邻域离群阈值μ, 调整参数λ

输出:离群点集O_S

(1) $O_S\leftarrow\varnothing$

(2) $\forall c_j$计算邻域关系矩阵$\boldsymbol{M}(N_{\{C_j\}})$

(3) for $j\leftarrow 1$ to m do

(4) 计算邻域关系矩阵 $\boldsymbol{M}(N_{C-\{C_j\}})$

(5) 计算邻域近似精度 $\alpha_{C-\{C_j\}}([x_i]_{C_j})$

$\backslash\backslash E=C-B=C-\{C_j\}=\{c_{e_1}, c_{e_2}, \cdots, c_{e_{m-1}}\}$

(6) for $k\leftarrow 1$ to $m-1$ do

(7) 计算邻域关系矩阵 $\boldsymbol{M}(N_{C-\{C_j, C_{e_k}\}})$

(8) 计算邻域近似精度 $\alpha_{C-\{C_j, C_{e_k}\}}([x_i]_{C_j})$

(9) end for

(10) 计算邻域粒离群度 $\mathrm{NGOD}\left(\delta_{\{C_j\}}(x_i)\right)$

(11) 计算权重 $W_{\{C_j\}}(x_i)$

(12) end for

(13) for $j\leftarrow 1$ to n do

(14) 计算邻域近似精度离群因子 $\mathrm{NAAOF}(x_i)$

(15) if $\mathrm{NAAOF}(x_i)>\mu$ then

(16) $O_S\leftarrow O_S\cup\{x_i\}$

(17) end if

(18) end for

算法复杂度:第(2)步的频度为$(n\times n)$, 对第(3~12)步的频度为$m\times(m+1+(m-1)\times n)$, 第(13~18)步的频度为$n$。故整体的频度为$(n^2+m^2\times n+m^2+m+n-m\times n)$。因此, 算法的时间复杂度为$O(m^2n)$, 空间复杂度为$O(mn)$。

3 数据实验及结果

实验以文献[14,18]预处理方法为基础, 并与适于标称的基于粒计算和粗糙集的离群检测(Outlier detection based on granular computing and rough set theory, ODGrCR)[14]、基于粒计算(Granular computing, GrC)的方法[10]、基于粗糙隶属度函数(Rough membership function, RMF)的方法[12]、适于数值属性的基于距离(Distance, DIS)的方法[5]以及适于混合属性的基于邻域信息熵的离群检测(Neighborhood information entropy-based outlier detection, NIEOD)[19]和基于邻域值差异度量的离群检测(Neighborhood value different metric-based outlier detection, NVDMOD)[20]算法进行对比, 以验证NAAOD算法的有效性。最后,从文献[18]中选择了6个离群点检测数据集(含标称、数值和混合属性)。这些数据集分别为Cred、Germ、Heart、Lymp、Wbc和Yeast。进一步,它们被分别导入信息系统I_{SC}、I_{SG}、I_{SH}、I_{SL}、I_{SW}和I_{SY}。为便于比较,同文献[14]的策略,分别从离群点数据集中选取2个子集,最终实验数据子集的基本特征信息描述如表1所示。

以离群点覆盖率(Coverage ratio, CR)[19,23]来评价所提算法。设$O_{S_{top_k}}(X)$是X中离群值排前k的对象, $O_{S_{true}}(X)$是真实离群点, 则$\left|O_{S_{top_k}}(X)\cap O_{S_{true}}(X)\right|$即检测出的真实离群点数。易知, $\mathrm{CR}(k)=|O_{S_{top_k}}(X)\cap O_{S_{true}}(X)|/|O_{S_{true}}(X)|$为离群点覆盖率。显然, TR($k$)不变时, CR($k$)越大算法性能越好。在实验中,取$k=\left|O_{S_{true}}(X)\right|$时的离群点覆盖率作为最终对比结果。

对于NAAOD、NIEOD和NVDMOD算法,其参数调节范围均为[0.1,2]且步长为0.1。最终,最优覆盖率作为实验结果。GrC算法的粒距离采用重叠距离法度量[10]。在粗糙集方法中, Lymphography和Wisconsin Breast Cancer对象的属性值均被当作标称类型[8]。除Yeast数据集,用MDL[24]离散化数

表 1 实验数据集描述

Table 1 Description of experimental data set

序号	数据子集	选取条件	条件属性数		对象数	离群点数
			数值	符号		
1	C_1	$C_1=\{x\in U_C\|f_C(x,C_1)='b'\}$	6	9	296	25
2	C_2	$C_2=\{x\in U_C\|f_C(x,C_{12})=2\}$			229	16
3	G_1	$G_1=\{x\in U_G\|f_G(x,C_{18})=1\}$	7	13	600	9
4	G_2	$G_2=\{x\in U_G\|f_G(x,C_{15})='\mathrm{A152}'\}$			531	4
5	H_1	$H_1=\{x\in U_H\|f_H(x,C_6)=1\}$	6	7	137	10
6	H_2	$H_2=\{x\in U_H\|f_H(x,C_9)=1\}$			131	4
7	L_1	$L_1=\{x\in U_L\|f_L(x,C_2)='\mathrm{no}'\vee f_L(x,C_8)='\mathrm{no}'\}$	0	18	90	5
8	L_2	$L_2=\{x\in U_L\|f_L(x,C_{13})='\mathrm{vesicles}'\vee f_L(x,C_{14})='\mathrm{no}'\}$			105	5
9	W_1	$W_1=\{x\in U_W\|f_W(x,C_4)=2\}$	9	0	42	5
10	W_2	$W_2=\{x\in U_W\|f_W(x,C_9)=1\}$			454	23
11	Y_1	$Y_1=\{x\in U_Y\|f_Y(x,C_2)\in[0.52,1]\}$	8	0	402	5
12	Y_2	$Y_2=\{x\in U_Y\|f_Y(x,C_4)=[0,0.4]\}$			948	5

值属性外,其他数值属性均使用Weka中等宽的离散化方法,其中离散化区间数为3。DIS算法采用欧氏距离度量和最小-最大归一化预处理。

NAAOD算法与对比算法在代表子集上的实验对比结果如表2所示。从表2中可以看出,NAAOD算法在6个混合属性数据子集上实现了较优的结果。例如,在C_1上,NAAOD算法的覆盖率为88.00%,与NIEOD和NVDMOD算法相等。而对于ODGrCR、GrC、RMF和DIS算法,其覆盖率分别为60.00%、28.00%、64.00%、68.00%,均小于NAAOD算法的覆盖率。因此,本文所提方法适用于混合属性数据的离群点检测。

此外,NAAOD算法在两个标称属性数据上均取得了最高的覆盖率。表2显示,NAAOD算法

表 2 对比实验结果

Table 2 Results of comparative experiments %

数据子集	NAAOD	NIEOD	NVDMOD	ODGrCR	GrC	RMF	DIS
C_1	88.00	88.00	88.00	60.00	28.00	64.00	68.00
C_2	87.50	87.50	87.50	75.00	12.50	75.00	56.25
G_1	33.33	33.33	33.33	22.22	22.22	22.22	33.33
G_2	50.00	50.00	50.00	25.00	50.00	25.00	25.00
H_1	90.00	90.00	90.00	80.00	70.00	90.00	80.00
H_2	75.00	75.00	75.00	50.00	50.00	50.00	50.00
L_1	100.00	80.00	60.00	100.00	80.00	60.00	80.00
L_2	100.00	80.00	60.00	80.00	80.00	80.00	80.00
W_1	100.00	100.00	100.00	80.00	80.00	80.00	80.00
W_2	91.30	86.96	86.96	78.26	73.91	78.26	78.26
Y_1	80.00	80.00	100.00	100.00	80.00	60.00	60.00
Y_2	100.00	100.00	100.00	80.00	60.00	60.00	60.00

在两个标称属性数据上的覆盖率为100%，即它能检测出全部离群点。然而，对于其他方法的覆盖率均小于或等于100%。同理，通过表2可以看出NAAOD算法在大部分数值属性数据的覆盖率均大于或等于其他对比算法。通过上述分析表明NAAOD算法也能有效处理标称和数值属性数据集。

阈值λ在NAAOD算法中起着重要的作用。C_1和W_2数据集被选取来探索参数对实验结果的敏感性。在C_1和W_2数据集上，其覆盖率随参数λ的变化曲线如图1所示。从图1可以看出在大部分数据集上随λ增加，其覆盖率先增加，然后呈现出趋于平衡的趋势。同时，也可以看到对于不同的数据集可以在多个参数λ值下取得最优值。

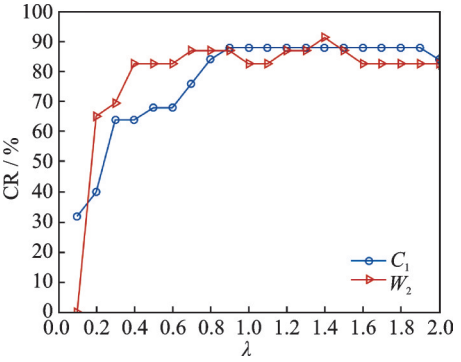

图1 在C_1和W_2数据集上覆盖率随参数λ的变化曲线

Fig.1 Variation curves of coverage rates with parameters λ on C_1 and W_2 datasets

4 结束语

基于邻域粗糙集框架，围绕混合属性离群检测问题，提出了基于邻域近似精度的具有自适应特性的离群检测方法。其离群因子融入了具有自适应特性的HNRM度量，且数值型属性无需离散化，能处理混合型数据集。数据实验结果表明，NAAOD算法能有效处理数值、标称和混合型属性数据集。接下来将从三支决策角度进一步研究基于粗糙集模型的其他相关离群点检测方法。

参考文献：

[1] HAWKINS D M. Identification of outliers[M]. London: Chapman and Hall, 1980.

[2] AGGARWAL C C. Outlier analysis[M]. Berlin: Springer, 2016.

[3] PICKANDS J. Statistical inference using extreme order statistics[J]. Annals of Statistics, 1975, 3(1): 119-131.

[4] JOHNSON T, KWOK I, NG R T. Fast computation of 2-dimensional depth contours[C]//Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining. Palo Alto: AAAI, 1998: 224-228.

[5] KNORR E M, NG R T. Algorithms for mining distance-based outliers in large datasets[C]//Proceedings of the 24th International Conference on Very Large Data Bases. San Francisco: Morgan Kaufmann, 1998: 392-403.

[6] RAMASWAMY S, RASTOGI R, SHIM K. Efficient algorithms for mining outliers from large data sets[C]//Proceedings of ACM Sigmod Record. [S.l.]: ACM, 2000: 427-438.

[7] BREUNIG M M, KRIEGEL H P, NG R T, et al. LOF: Identifying density-based local outliers[C]//Proceedings of ACM Sigmod Record. [S.l.]: ACM, 2000: 93-104.

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

[9] PAWLAK Z. Rough sets[J]. International Journal of Computer & Information Sciences, 1982, 11(5): 341-356.

[10] CHEN Y M, MIAO D, WANG R. Outlier detection based on granular computing[C]//Proceedings of International Conference on Rough Sets and Current Trends in Computing. Berlin, Germany: Springer, 2008: 283-292.

[11] JIANG F, SUI Y, CAO C. Outlier detection using rough set theory[C]//Proceedings of International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing. Berlin, Germany: Springer, 2005: 79-87.

[12] JIANG F, SUI Y, CAO C. A rough set approach to outlier detection[J]. International Journal of General Systems, 2008, 37(5): 519-536.

[13] JIANG F, SUI Y, CAO C. A hybrid approach to outlier detection based on boundary region[J]. Pattern Recognition Letters,

2011, 32(14): 1860-1870.

[14] JIANG F, CHEN Y M. Outlier detection based on granular computing and rough set theory[J]. Applied Intelligence, 2015, 42(2): 303-322.

[15] HU Q H, YU D, LIU J, et al. Neighborhood rough set based heterogeneous feature subset selection[J]. Information Sciences, 2008, 178(18): 3577-3594.

[16] CHEN Y, XUE Y, MA Y, et al. Measures of uncertainty for neighborhood rough sets[J]. Knowledge-Based Systems, 2017, 120: 226-235.

[17] CHEN Y M, MIAO D Q, ZHANG H Y. Neighborhood outlier detection[J]. Expert Systems with Applications, 2010, 37(12): 8745-8749.

[18] YUAN Z, CHEN H M, LI T R, et al. Fuzzy information entropy-based adaptive approach for hybrid feature outlier detection [J]. Fuzzy Sets and Systems, 2021, 421: 1-28.

[19] YUAN Z, ZHANG X Y, FENG S. Hybrid data-driven outlier detection based on neighborhood information entropy and its developmental measures[J]. Expert Systems with Applications, 2018, 112: 243-257.

[20] 袁钟, 冯山. 基于邻域值差异度量的离群点检测算法[J]. 计算机应用, 2018, 7: 1905-1909.
YUAN Zhong, FENG Shan. Outlier detection algorithm based on neighborhood value difference metric[J]. Journal of Compuer Application, 2018, 7: 1905-1909.

[21] 袁钟, 张贤勇, 冯山. 邻域粗糙集中基于序列的混合型属性离群点检测[J]. 小型微型计算机系统, 2018, 39(6): 1317-1322.
YUAN Zhong, ZHANG Xianyong, FENG Shan. Sequence-based mixed attribute outlier detection in neighborhood rough sets [J]. Journal of Chinese Computer Systems, 2018, 39(6): 1317-1322.

[22] WILSON D R, MARTINEZ T R. Improved heterogeneous distance functions[J]. Journal of Artificial Intelligence Research, 1997, 6: 1-34.

[23] AGGARWAL C C, YU P S. Outlier detection for high dimensional data[J]. ACM Sigmod Record, 2001, 30(2): 37-46.

[24] FAYYAD U, IRANI K. Multi-interval discretization of continuous-valued attributes for classification learning[C]// Proceedings of the 13th International Conference on Artificial Intelligence. New York: ACM, 1993: 1022-1027.

作者简介:

张玉婷(1993-),女,硕士研究生,研究方向:粗糙集、数据挖掘,E-mail: zhangyutingthemail@foxmail.com。

冯山(1967-),**通信作者**,男,教授,硕士生导师,研究方向:粗糙集、数据挖掘,E-mail: fengshanrq@sohu.com。

(编辑:陈珺)