



文章编号:1001-9081(2018)07-1905-05

DOI:10.11772/j.issn.1001-9081.2017123028

基于邻域值差异度量的离群点检测算法

袁 钟, 冯 山^{*}

(四川师范大学 数学与软件科学学院, 成都 610068)

(* 通信作者电子邮箱 fengshanrq@sohu.com)

摘要:针对离群点检测中传统距离法不能有效处理符号型属性和经典粗糙集方法不能有效处理数值型属性的问题,利用邻域粗糙集的粒化特征提出了改进的邻域值差异度量(NVDM)方法进行离群点检测。首先,将属性取值归一化并以混合欧氏重叠度量(HEOM)和具有自适应特征的邻域半径构建邻域信息系统(NIS);其次,以NVDM构造对象的邻域离群因子(NOF);最后,设计并实现了基于邻域值差异度量的离群点检测(NVDMOD)算法,该算法在计算单属性邻域覆盖(SANC)的方式上充分利用有序二分和近邻搜索思想改进了传统的无序逐一计算模式。在UCI标准数据集上与现有离群点检测算法——邻域离群点检测(NED)算法、基于距离的离群点检测(DIS)算法和K最近邻(KNN)算法进行了实验对比、分析。实验结果表明,NVDMOD算法具有更好的适应性和有效性,为混合型属性数据集的离群点检测提供了一条更有效的新途径。

关键词:离群点检测;邻域粗糙集;邻域值差异度量;混合型属性;数据挖掘

中图分类号: TP274 **文献标志码:**A

Outlier detection algorithm based on neighborhood value difference metric

YUAN Zhong, FENG Shan^{*}

(College of Mathematics and Software Science, Sichuan Normal University, Chengdu Sichuan 610068, China)

Abstract: Aiming at the problems that symbolic attribute data set can not be processed effectively with traditional distance measure method and numerical attribute data set can not be processed effectively by classical rough set method, an improved method of Neighborhood Value Difference Metric (NVDM) was proposed for outlier detection by utilizing the granulation features of neighborhood rough set. Firstly, with attribute values being normalized, the Neighborhood Information System (NIS) was constructed based on optimized Heterogeneous Euclidian-Overlap Metric (HEOM) and neighborhood radius with adaptive characteristic. Secondly, Neighborhood Outlier Factor (NOF) of data object was constructed based on the NVDM. Finally, a Neighborhood Value Difference Metric-based Outlier Detection (NVDMOD) algorithm was designed and implemented, which improves the traditional unordered one by one model via making full use of the idea of ordered binary and nearest neighbor search in computing Single Attribute Neighborhood Cover (SANC). The NVDMOD algorithm was analyzed and compared with existing outlier detection algorithms including NEighborhood outlier Detection (NED) algorithm, DIstance-based outlier detection (DIS) algorithm and K-Nearest Neighbor (KNN) algorithm on UCI standard data sets. The experimental results show that NVDMOD algorithm has much higher adaptability and effectiveness, and it provides a more effective new method for outlier detection of mixed attribute data sets.

Key words: outlier detection; neighborhood rough set; Neighborhood Value Difference Metric (NVDM); mixed attribute; data mining

0 引言

离群点是数据集中数据对象特征显著区别于其他数据对象的数据对象^[1],其出现往往隐含或预示具有特殊意义的事件或现象发生。在入侵与欺诈检测、医疗处理、公共安全等领域,离群点检测技术具有十分重要的应用^[2-4]。

离群点检测研究最早出现于统计学领域^[5]。后来,Knorr等^[6-7]将其引入到数据挖掘领域。现有离群点检测方法主要有三类:1) 基于统计^[5];2) 基于邻近性^[6-8];3) 基于聚类^[9]。其中,基于邻近性的离群点检测有基于距离、基于网格和基于

密度三种方式。

符号型属性值是离散的,用传统距离法检测符号型属性数据集的离群点效果并不理想。为此,人们引入了粗糙集下的符号型属性离群点检测^[10],但它基于等价关系和等价类建模,用于处理数值型属性数据集时要先对属性值离散化,既增加了处理时间,还带来了明显的数据信息损失,影响检测精度。

为解决数值型属性的粗糙计算问题,结合邻域概念和邻域关系,文献[11]建立了邻域粗糙集模型,它是属性约简、特征选择、分类和不确定性推理研究中的重要工具^[12];但是,利

收稿日期:2017-12-25;修回日期:2018-02-07;录用日期:2018-02-26。 基金项目:国家自然科学基金资助项目(61673285);四川省青年科技基金资助项目(2017JQ0046);四川省教育厅自然科学重点基金资助项目(15ZB0029)。

作者简介:袁钟(1991—),男,四川井研人,硕士研究生,主要研究方向:粗糙集、数据挖掘; 冯山(1967—),男,重庆丰都人,教授,博士,主要研究方向:粗糙集、数据挖掘。



用邻域粗糙集模型进行离群点检测的研究并不多见,文献[13]进行了这方面的研究尝试,但是其邻域半径直接给定,具有较强的主观性和随机性。

针对离群点检测中传统距离法不能有效处理符号型属性和经典粗糙集方法不能有效处理数值型属性的问题,本文以邻域粗糙集模型为基础,提出了改进的邻域值差异度量(Neighborhood Value Difference Metric, NVDM)进行离群点检测。该方法通过定义邻域值差异度量来构造表征对象离群程度的邻域离群因子(Neighborhood Outlier Factor, NOF),从而,设计并实现了基于邻域值差异度量的离群点检测(Neighborhood Value Difference Metric-based Outlier Detection, NVMOD)算法。所提算法拓展了离群点检测的传统距离法和粗糙集法,在计算单属性邻域覆盖(Single Attribute Neighborhood Cover, SANC)时改进了传统的无序逐一计算模式,使得算法时间复杂度降低到了对数级,明显优于现有传统无序逐一比较模式。UCI标准数据集实验表明,改进算法能有效分析和处理符号型、数值型和混合型属性数据集的离群点检测问题,且适应能力更强。

1 邻域粗糙集

设 $IS = (U, A, V, f)$ 是信息系统, $U = \{x_1, x_2, \dots, x_n\}$ 是非空有限数据对象集, A 是数据对象的非空有限属性集, $V = \bigcup_{a \in A} V_a$ 是所有对象的属性值域 V_a 的并, 映射函数 $f: U \times A \rightarrow V$, $\forall x \in U$ 以及 $a \in A$, 有 $f(x, a) \in V_a$ 。当 A 由条件属性集 $C = \{c_1, c_2, \dots, c_m\}$ 和决策属性集 D 构成时, IS 称为决策系统^[12], 简记 $DS = (U, C \cup D, V, f)$ 。

对 $\forall x, y, z \in U$ 和 $B \subseteq C$, 关联于属性子集 B 的距离函数 $d_B: U \times U \rightarrow \mathbf{R}^+$ (\mathbf{R}^+ 非负实数集) 应满足以下条件:

- 1) $d_B(x, y) \geq 0, d_B(x, x) = 0$ (非负性);
- 2) $d_B(x, y) = d_B(y, x)$ (对称性);
- 3) $d_B(x, z) \leq d_B(x, y) + d_B(y, z)$ (三角式)。

如果 $B = \{c_{j_1}, c_{j_2}, \dots, c_{j_k}\} \subseteq C$ ($1 \leq k \leq m$), 距离函数 d_B 的闵可夫斯基距离计算公式如下:

$$d_B^p(x, y) = \sqrt[p]{\sum_{h=1}^k |f(x, c_{j_h}) - f(y, c_{j_h})|^p}$$

因等价关系和等价类只能处理符号型属性, 可将距离函数和邻域半径 ε 结合并用来对论域 U 中的数据对象粒化, 形成具有相似性特征的邻域关系和邻域类, 进而构建可同时处理符号型属性和数值型属性的邻域信息系统(Neighborhood Information System, NIS)。

定义1 邻域。对 $\forall x \in U, B \subseteq C$ 和 $\varepsilon \geq 0$, x 在属性集 B 上的 ε -邻域定义为:

$$n_B^\varepsilon(x) = \{y \mid d_B(x, y) \leq \varepsilon, y \in U\} \quad (1)$$

定义2 邻域关系。对 $\forall B \subseteq C$ 和 $\varepsilon \geq 0$, 称 $nr_B^\varepsilon = \{(x_i, x_j) \mid x_i, x_j \in U \text{ 且 } x_j \in n_B^\varepsilon(x_i)\}$ 为 U 上的 B - ε 邻域关系。

显然, 邻域关系是相似关系, 它刻画了对象间的距离相似性或不可区分性。实际上: nr_B^0 是 U 上最细等价关系, 适于处理符号型属性; $nr_B^\varepsilon (\varepsilon \neq 0)$ 是 U 上最粗等价关系, 适于处理数值型属性。它们是相似关系的特殊情形。

定义3 邻域覆盖、邻域类和邻域知识。对 $\forall B \subseteq C$ 和

$\varepsilon \geq 0, U/nr_B^\varepsilon$ 构成 U 的一个邻域覆盖, 它对应一个邻域类, 也称为 U 上的一个邻域知识。

定义4 邻域信息系统。假设 nr_B^ε 是 U 上的 B - ε 邻域关系, $NR_C^\varepsilon = \{nr_B^\varepsilon : B \subseteq C\}$ 是 U 上的全体邻域关系, 则 $NIS = (U, NR_C^\varepsilon, V, f)$ 是关于 IS 的邻域信息系统, 相应地, $NDS = (U, NR_C^\varepsilon \cup D, V, f)$ 称为邻域决策系统。

以邻域信息系统为基础, 可以利用邻域粗糙性引起的邻域间的值差异度量对论域中对象之间的差异性或离群性进行度量, 从而发现离群点。

2 基于邻域值差异度量的离群点检测

2.1 基于邻域值差异度量的离群点检测方法

用邻域粗糙集进行数据处理时通常会存在数据描述的量级和量纲差异。为使不同数据类型的表述都得到有效的数据处理结果, 需要对数值型属性值进行归一化、标准化处理。本文采用的最小—最大法归一化的计算公式如下:

$$F(f(x_i, c_j)) = \frac{f(x_i, c_j) - \min_{c_j}}{\max_{c_j} - \min_{c_j}} \quad (2)$$

其中: \max_{c_j} 和 \min_{c_j} 为 U 中对象关于属性 c_j 的最大取值和最小取值。显然, 标准化属性取值区间为 $[0, 1]$ 。

为同时有效度量数值型和混合型属性值的差异, 可用混合欧氏重叠度量(Heterogeneous Euclidian-Overlap Metric, HEOM)^[14] 进行邻域距离度量。

定义5 混合欧氏重叠度量。对 $\forall x, y \in U, x$ 和 y 的混合欧氏重叠度量 $HEOM_B(x, y)$ 定义为:

$$HEOM_B(x, y) = \sqrt{\sum_{h=1}^k d_{c_{j_h}}(x, y)^2} \quad (3)$$

其中:

$$d_{c_{j_h}}(x, y) = \begin{cases} 0, & \text{如果 } c_{j_h} \text{ 是符号型属性且 } f(x, c_{j_h}) = f(y, c_{j_h}) \\ 1, & \text{如果 } c_{j_h} \text{ 是符号型属性且 } f(x, c_{j_h}) \neq f(y, c_{j_h}) \\ |f(x, c_{j_h}) - f(y, c_{j_h})|, & \text{如果 } c_{j_h} \text{ 是数值型属性} \end{cases}$$

对象邻域半径的设定一般由专家根据经验确定, 具有较强的主观性和随机性。减少邻域构造判定算法对专家所定参数的敏感度, 是提升离群点检测算法准确率的客观基础。能够将数据对象的属性取值分布信息和专家知识融合的邻域半径参数确定法更加合理、有效。为此, 本文提出了 x 在属性 c_j 上的邻域半径设置新方法, 它具有很好的自适应特征:

$$\varepsilon_{c_j} = \begin{cases} 0, & \text{如果属性 } c_j \text{ 为符号型属性} \\ std(c_j)/\lambda, & \text{如果属性 } c_j \text{ 为数值型属性} \end{cases} \quad (4)$$

其中: $std(c_j)$ 是 c_j 属性的取值标准差, 用于衡量 c_j 的均值分散程度, $std(c_j)$ 大表示大部分数据对象在 c_j 上的取值与其均值间的差异大, $std(c_j)$ 小表示数据对象在 c_j 上的取值与均值接近。 λ 是专家预设的用于调整邻域半径大小的参数, $\lambda < 1$ 时邻域半径应大于属性值标准差; $\lambda = 1$ 时邻域半径即属性值标准差; $\lambda > 1$ 时邻域半径应小于属性值标准差。这种主、客观融合的邻域半径调节法兼顾了专家经验和属性取值分布特征对对象离群性的影响, 为有效的自适应离群点检测奠定了基础。

通过对象邻域及其距离度量可刻画对象间的相似性或不



可区分性。值差异度量 (Value Difference Metric, VDM)^[15] 是度量符号型属性距离的一种有效的非加权函数。假定 x 和 y 是 U 中对象, F 是 U 的特征集, x_f 和 y_f 是 x 和 y 在 f 上的取值, $d_f(x_f, y_f)$ 是 x_f 和 y_f 的距离。对象 x 和 y 的 VDM 可定义如下:

$$VDM(x, y) = \sum_{f \in F} d_f(x_f, y_f)$$

其中: $d_f(x_f, y_f) = (P(x_f) - P(y_f))^2$, $P(x_f)$ 和 $P(y_f)$ 是 x 和 y 在 f 上的取值概率。

以此类推, 为度量数值型属性取值的离群程度, 可通过对对象邻域概念将给定对象的邻域半径内的对象集成, 进而对数值型属性取值进行离群度量。为此, 可定义邻域值差异度量 (NVDM)、邻域离群因子 (NOF) 及邻域离群点概念如下。

定义 6 邻域值差异度量。给定 $\lambda > 0$, 对 $\forall x_i, x_j \in U, x_i, x_j$ 的邻域值差异度量 $NVDM(x_i, x_j)$ 为:

$$NVDM(x_i, x_j) = \sum_{c \in C} d_c(x_i, x_j) \quad (5)$$

其中: $d_c(x_i, x_j) = (|n_{\{c\}}^{e_c}(x_i)| / |U| - |n_{\{c\}}^{e_c}(x_j)| / |U|)^2$; $|\cdot|$ 是集合 \cdot 的势。

实际上, 对象离群程度可用邻域离群因子度量, 可将文献 [13] 的邻域离群因子开平方以加大对象离群程度的变化对其离群特性的正面影响。

定义 7 邻域离群因子。对 $\forall x_i \in U, x_i$ 的邻域离群因子的计算公式如下:

$$NOF(x_i) = \sum_{j=1, j \neq i}^n \sqrt{NVDM(x_i, x_j)} \quad (6)$$

定义 8 基于邻域值差异度量的离群点。令 μ 为给定的离群点判定阈值, 对 $\forall x \in U$, 如果 $NOF(x) > \mu$, 称 x 为 U 中基于邻域值差异度量的离群点。

2.2 基于邻域值差异度量的离群点检测算法

基于邻域和值差异度量概念, 本文提出了基于邻域值差异度量的离群点检测 (NVDMOD) 算法 (算法 2)。它在单属性邻域覆盖 (SANC) (算法 1) 计算时采取有序二分近邻搜索模式, 效率较传统无序逐一比较模式有显著提升。在数据结构设计上, 新算法用数组首行存放 U 中对象按属性 c_j 升序排列的结果, 数组第二行存放对象在 U 中的原始顺序号。

算法 1 单属性邻域覆盖 (SANC) 算法。

输入: $IS = (U, C, V, f)$, λ 和 j , 其中 $|U| = n$ 。
输出: 单属性邻域覆盖 $U/nr_{\{c_j\}}$ 。

- 1) $U/nr_{\{c_j\}} \leftarrow \emptyset$ /* 初始化 */
- 2) $N \leftarrow \emptyset$
- 3) $RankIndex[1..n][1..2] \leftarrow Ascend_sort(U)$ /* U 中对象按 c_j 升序排列 */
- 4) for $i = 1$ to n do
- 5) $k \leftarrow i$
- 6) while $k > 0$ do
- 7) if $HEOM_{\{c_j\}}(RankIndex[i][1], RankIndex[k][1]) \leq \varepsilon_{c_j}$,
- 8) $k \leftarrow k - 1$
- 9) else
- 10) break
- 11) end if
- 12) end while
- 13) $a \leftarrow k + 1$ /* a : 对象邻域的下限顺序号 */
- 14) $k \leftarrow i + 1$
- 15) while $k < n$ do

```

16)   if  $HEOM_{\{c_j\}}(RankIndex[i][1], RankIndex[k][1]) \leq \varepsilon_{c_j}$ ,
17)      $k \leftarrow k + 1$ 
18)   else
19)     break
20)   end if
21) end while
22)  $b \leftarrow k - 1$  /*  $b$ : 对象邻域的上限顺序号 */
23)  $N \leftarrow RankIndex[a..b][1]$ 
24)  $U/nr_{\{c_j\}}[(RankIndex[i][2])] \leftarrow N$ 
25) /* 将邻域  $N$  存入  $U/nr_{\{c_j\}}$  的第  $RankIndex[i][2]$  行 */
26) end for
27) return  $U/nr_{\{c_j\}}$ 

```

SANC 算法第 3) 步采用堆排序^[16], 第 4) ~ 25) 步的频度为 n , 第 6) ~ 12) 及 15) ~ 21) 步的频度为 n , 理论时间复杂度 $O(n^2)$ 与传统无序逐一比较算法相同, 但由于 SANC 算法第 3) 步进行了属性值的预排序, 计算对象的单属性邻域时可以利用该有序性进行二分近邻搜索。由于邻域中的对象都分布在邻近位置, 对给定对象的邻域搜索比较次数会比 n 小很多, 故 SANC 算法的实际计算复杂度远低于 $O(n^2)$ 。假定对象属性的取值等概率均匀分布, 此时 ε 小, 有序二分近邻搜索的复杂度将接近 $O(n)$ 。综上, SANC 算法的实际时间复杂度为 $O(n \log n)$, 明显低于传统的无序逐一比较算法。

算法 2 基于邻域值差异度量的离群点检测 (NVDMOD) 算法

输入: $IS = (U, C, V, f)$, 阈值 μ, λ , 其中 $|U| = n, |C| = m$ 。
输出: 基于邻域值差异度量的离群点集 OS (Outlier Set)。

- 1) $OS \leftarrow \emptyset$
- 2) for $j \leftarrow 1$ to m do
- 3) 计算 $U/nr_{\{c_j\}}$ /* 由 SANC 算法计算 */
- 4) end for
- 5) for $i \leftarrow 1$ to n do
- 6) for $j \leftarrow 1$ to n do
- 7) for $k \leftarrow 1$ to m do
- 8) if $i \neq j$
- 9) 计算 $d_{\{c_k\}}(x_i, x_j)$ /* x_i, x_j 的单属性距离 */
- 10) end if
- 11) end for
- 12) end for
- 13) 计算 $NVDM(x_i, x_j)$ /* x_i, x_j 的邻域值差异度量 */
- 14) 计算 $NOF(x_i)$ /* 对象 x_i 的邻域离群因子 */
- 15) if $NOF(x_i) > \mu$
- 16) $OS \leftarrow OS \cup \{x_i\}$
- 17) end if
- 18) end for
- 19) return OS

对于 NVDMOD 算法, 由于 SANC 算法的复杂度为 $O(n \log n)$, 它重复 m 次, 同时, 第 5) ~ 18) 步的频度为 $m \times n^2$, 因此, NVDMOD 算法的时间复杂度为 $O(mn^2)$, 空间复杂度为 $O(mn)$ 。

3 实验结果及其分析

本章对 NVDMOD 算法、邻域离群点检测 (NEighborhood outlier Detection, NED) 算法^[13] (邻域半径直接给定)、基于距



离的离群点检测(DIStance-based outlier detection, DIS)方法^[6](传统距离检测方法)和K最近邻(K-Nearest Neighbors, KNN)算法^[17](传统距离检测方法)的性能进行实验比较,验证NVDMOD算法的有效性和适应性。

3.1 实验环境

为测试NVDMOD算法的有效性,选取Annealing和Wisconsin Breast Cancer两个数据集进行实验。

首先从UCI机器学习库中获得上述数据集^[18]。实验平台配置:处理器Intel Core i5-2400;主频3.10 GHz;内存4 GB;操作系统Windows 7;编程环境Matlab R2015b。

为增强实验结果的可比性,本文采用文献[19]中的离群点检测方法评价体系以给定数据集上找出的离群点占比衡量算法的有效性。

3.2 Annealing数据集

Annealing数据集有798个数据对象、37个条件属性和1个决策属性。其中,条件属性包括30个符号型和7个数值型。数据对象分为5类,类3以外的数据对象均为离群点,共190个离群点。Annealing邻域信息系统记为NIS_A,邻域半径参数 $\lambda_A=0.2$ 。4种算法的实验结果如表1所示。

表1 邻域信息系统NIS_A上的实验结果

Tab. 1 Experimental results in NIS_A

| 离群程度前k%的对象 | 对象数 | 属于离群点的对象数 | | | | | 覆盖率/% | | |
|------------|-----|-----------|-----|-----|-----|--------|-------|-------|-------|
| | | NVD-MOD | NED | DIS | KNN | NVDMOD | NED | DIS | KNN |
| 10.03 | 80 | 64 | 51 | 33 | 21 | 33.68 | 26.84 | 17.37 | 11.05 |
| 13.16 | 105 | 75 | 67 | 44 | 30 | 39.47 | 35.26 | 23.16 | 15.79 |
| 17.54 | 140 | 81 | 81 | 61 | 41 | 42.63 | 42.63 | 32.11 | 21.58 |
| 21.93 | 175 | 85 | 84 | 77 | 58 | 44.74 | 44.21 | 40.53 | 30.33 |
| 26.19 | 209 | 99 | 92 | 84 | 62 | 52.11 | 48.42 | 44.21 | 32.63 |

其中:离群程度前k%的对象(对象数)是指将数据对象按离群程度值由高到低排序后,用于对比分析的对象子集比例(前k%)及其对象数;属于离群点的对象数是指离群度前k%的对象中属于离群点的对象数;覆盖率是指属于离群点对象数占离群点总数的比例。

Annealing数据集是混合型属性数据集。由表1可见,NVDMOD算法准确率明显高于其他3种算法,如离群程度前10.03%的80个数据对象中,它能检测出64个离群点,而NED、KNN和DIS算法只分别检测出了51、33和21个。由此可见,NVDMOD算法适用于混合型数据集,且优于其他算法。

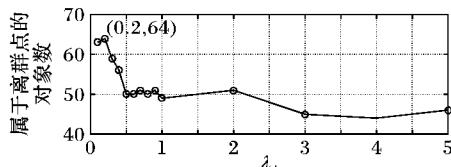


图1 离群点数随 λ_A 变化的折线图(离群程度前10.03%)

Fig. 1 Line chart of number of detected outliers change with λ_A (where the top outlier degree is 10.03%)

离群程度前10.03%的对象中,NVDMOD算法检测出的离群点数随 λ_A 变化如图1所示,当 $\lambda_A=0.2$ 时效果最好。 $0 < \lambda_A \leq 2$ 时,NVDMOD算法总体优于其他3种算法,而其余情形时优于DIS和KNN算法,接近NED算法。

3.3 Wisconsin Breast Cancer数据集

Wisconsin Breast Cancer数据集有699个对象,分成benign(65.5%)和malignant(34.5%)两类,对象的描述由9个数值型属性完成。为形成不均匀分布数据集,仿照Harkins等提出的方法^[20]移去了一些属于malignant类的对象,最终的数据集包含了483个对象,其中39个属于malignant类。

将malignant类作为离群点,数据集的邻域信息系统记为NIS_w,邻域半径参数 $\lambda_w=0.4$ 。在NIS_w上实验结果如表2所示。

表2 邻域信息系统NIS_w上的实验结果

Tab. 2 Experimental results in NIS_w

| 离群程度前k%的对象 | 对象数 | 属于离群点的对象数 | | | | | 覆盖率/% | | |
|------------|-----|-----------|-----|-----|-----|--------|--------|--------|--------|
| | | NVD-MOD | NED | DIS | KNN | NVDMOD | NED | DIS | KNN |
| 0.83 | 4 | 4 | 4 | 4 | 4 | 10.26 | 10.26 | 10.26 | 10.26 |
| 1.66 | 8 | 8 | 7 | 7 | 7 | 20.51 | 17.95 | 17.95 | 17.95 |
| 3.31 | 16 | 16 | 14 | 14 | 13 | 41.03 | 35.90 | 35.90 | 33.33 |
| 4.97 | 24 | 24 | 19 | 21 | 20 | 61.54 | 48.72 | 53.85 | 51.28 |
| 6.63 | 32 | 28 | 26 | 28 | 27 | 71.79 | 66.67 | 32.29 | 69.23 |
| 8.28 | 40 | 33 | 31 | 32 | 32 | 84.62 | 79.49 | 82.05 | 82.05 |
| 9.94 | 48 | 39 | 36 | 36 | 38 | 100.00 | 92.31 | 92.31 | 97.44 |
| 11.59 | 56 | 39 | 38 | 39 | 39 | 100.00 | 97.44 | 100.00 | 100.00 |
| 13.25 | 64 | 39 | 39 | 39 | 39 | 100.00 | 100.00 | 100.00 | 100.00 |

从表2可知,在Wisconsin Breast Cancer数据集中,NVDMOD算法具有最好性能,通过对离群程度前9.94%的48个数据对象进行检测,即可检测出所有离群点。该数据集的9个属性全为数值型属性。由此可见,NVDMOD算法处理数值型数据对象问题也能取得很好的效果。

离群程度前9.94%的48个数据对象中,NVDMOD算法检测出的离群点数随 λ_w 变化的规律如图2所示,当 $\lambda_w=0.4$ 时效果最好; $\lambda_w \geq 0.4$ 时NVDMOD算法大致具有稳定性和优化性。

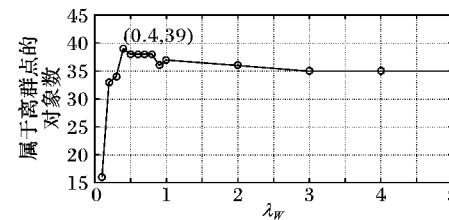


图2 离群点数随 λ_w 变化的折线图(离群程度前9.94%)

Fig. 2 Line chart of number of detected outliers change with λ_w (where the top outlier degree is 9.94%)

4 结语

针对传统距离法不能有效处理符号型属性和经典粗糙集方法不能有效处理数值型属性问题,通过归一化预处理、HEOM距离选取和具有自适应特征的邻域半径设定,本文构建了基于邻域关系和邻域类的邻域信息系统,利用邻近多数类的不确定性颗粒性质,以对象邻域值差异度量为基础进行离群点检测,较好地融合和拓展了传统距离法和粗糙集方法,能够更有效地处理符号型、数值型和混合型属性数据集。实验结果表明,所提算法在各类属性组合情形下都有更好的适应能力。本文的研究是针对邻域粗糙集方法及其应用的拓



展。后续研究中,可以考虑通过属性序列和属性集序列^[10]集成离群因子,以提高算法检测结果的有效性和算法效率。另外,还可以考虑将各类属性取值的离群特征信息融合到对象的离群度量模型中,进一步提高算法效率;以及引入统计检验进一步分析各算法的性能优劣。

参考文献 (References)

- [1] HAWKINS D. Identification of Outliers [M]. London: Chapman and Hall, 1980: 1–2.
- [2] 王习特,申德荣,白梅,等. BOD:一种高效的分布式离群点检测算法[J].计算机学报,2016,39(1):36–51.(WANG X T, SHEN D R, BAI M, et al. BOD: an efficient algorithm for distributed outlier detection [J]. Chinese Journal of Computers, 2016, 39(1): 36 –51).
- [3] 邹云峰,张忻,宋世渊,等.基于局部密度的快速离群点检测算法[J].计算机应用,2017,37(10):2932–2937.(ZOU Y F, ZHANG X, SONG S Y, et al. Fast outlier detection algorithm based on local density [J]. Journal of Computer Applications, 2017, 37 (10): 2932 –2937.)
- [4] HAN J W, KAMBER M, PEI J. Data Mining: Concepts and Techniques [M]. 3rd ed. San Francisco: Morgan Kaufmann, 2011: 543 –583.
- [5] ROUSSEEUW P J, LEROY A M. Robust Regression and Outlier Detection [M]. Hoboken: John Wiley and Sons, 1987: 1–18.
- [6] KNORR E M, NG R T, TUCAKOV V. Distance-based outliers: algorithms and applications [J]. The VLDB Journal, 2000, 8 (3): 237 –253.
- [7] KNORR E, NG R. A unified notion of outliers: properties and computation [C]// Proceedings of the 1997 International Conference on Knowledge Discovery & Data Mining. Menlo Park, CA: AAAI Press, 1997: 219 –222.
- [8] 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, 2000: 93 –104.
- [9] JAIN A K, MURTY M N, FLYNN P J. Data clustering: a review [J]. ACM Computing Surveys, 1999, 31(3): 264 –323.
- [10] JIANG F, SUI Y F, CAO C G. An information entropy-based approach to outlier detection in rough sets [J]. Expert Systems with Applications, 2010, 37(9): 6338 –6344.
- [11] LIN T Y. Neighborhood systems-application to qualitative fuzzy and rough sets [C]// Advances in Machine Intelligence and Soft-Computing. Durham: Department of Electrical Engineering, 1997: 132 –155.
- [12] HU Q H, YU D R, LIU J F, et al. Neighborhood rough set based heterogeneous feature subset selection [J]. Information Sciences, 2008, 178(18): 3577 –3594.
- [13] CHEN Y M, MIAO D Q, ZHANG H Y. Neighborhood outlier detection [J]. Expert Systems with Applications, 2010, 37 (12): 8745 –8749.
- [14] WILSON D R, MARTINEZ T R. Improved heterogeneous distance functions [J]. Journal of Artificial Intelligence Research, 1997, 6 (1): 1 –34.
- [15] STANFILL C, WALTZ D. Toward memory-based reasoning [J]. Communications of the ACM, 1986, 29(12): 1213 –1228.
- [16] WILLIAMS J W J. Algorithm 232 (heapsort) [J]. Communications of the ACM, 1964, 7(6): 347 –348.
- [17] RAMASWAMY S, RASTOGI R, SHIM K. Efficient algorithms for mining outliers from large datasets [C]// Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2000: 427 –438.
- [18] BAY S D. The UCI KDD repository [EB/OL]. [2017-05-12]. <http://kdd.ics.uci.edu>.
- [19] AGGARWAL C C, SINGH P S. Outlier detection for high dimensional data [C]// Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2001: 37 –46.
- [20] HARKINS, HE H X, WILLIAMS G J, et al. Outlier detection using replicator neural networks [C]// Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery. Berlin: Springer, 2002: 170 –180.

This work is partially supported by the National Natural Science Foundation of China (61673258), the Sichuan Youth Science and Technology Foundation (2017JQ0046), the Scientific Research Project of Sichuan Provincial Education Department (15ZB0029).

YUAN Zhong, born in 1991, M. S. candidate. His research interests include rough set, data mining.

FENG Shan, born in 1967, Ph. D., professor. His research interests include rough set, data mining.

25(2): 292 –296.

This work is partially supported by the National Natural Science Foundation of China (61371090).

GUAN Haoyuan, born in 1993, M. S. candidate. His research interests include intelligent information processing.

ZHU Bin, born in 1969, M. S., associate professor. His research interests include intelligent information processing.

LI Guanyu, born in 1963, Ph. D., professor. His research interests include intelligent information processing.

ZHAO Ling, born in 1993, M. S. candidate. Her research interests include intelligent information processing.

(上接第1904页)

- [16] RIVERO C R, JAMIL H M. Efficient and scalable labeled subgraph matching using SGMatch [J]. Knowledge and Information Systems, 2017, 51(1): 61 –87.
- [17] HE H, SINGH A K. Graphs-at-a-time: query language and access methods for graph databases [C]// SIGMOD 2008: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2008: 405 –418.
- [18] HOFFART J, SUCHANEK F M, BERBERICH K, et al. YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia [J]. Artificial Intelligence, 2013, 194: 28 –61.
- [19] FELLBAUM C, MILLER G. WordNet: an electronic lexical database [J]. Library Quarterly Information Community Policy, 1998,