文章编号: 1001-0920 (2013) 02-0188-05

粗糙集中的距离度量与离群点检测

江 峰[1], 眭跃飞[2], 曹存根[2]

(1. 青岛科技大学 信息科学技术学院，山东 青岛 266061；2. 中国科学院 计算技术研究所，北京 100080)

摘 要: 针对传统的基于距离的离群点检测方法不能有效地处理具有离散型属性数据集的问题, 将基于距离的离群点检测方法引入粗糙集理论, 利用粗糙集解决离散型属性的处理问题. 首先, 在粗糙集的框架中提出 3 种面向离散型属性的距离度量; 然后, 针对这 3 种距离度量分别设计出相应的离群点检测算法, 用于从包含离散型属性的数据集中检测离群点; 最后, 通过在 2 个包含离散型属性的 UCI 数据集上的实验, 验证了这些算法的可行性和有效性.

关键词: 粗糙集理论；离群点检测；数据挖掘；距离度量；离散型属性

中图分类号: TP311 **文献标志码:** A

Distance metrics and outlier detection in rough sets

JIANG Feng[1], *SUI Yue-fei*[2], *CAO Cun-gen*[2]

(1. College of Information Science and Technology，Qingdao University of Science and Technology，Qingdao 266061, China；2. Institute of Computing Technology，Chinese Academy of Sciences，Beijing 100080，China．Correspondent: JIANG Feng，E-mail：jiangkong@163.net)

Abstract: The traditional distance-based outlier detection method can not effectively deal with the data sets containing discrete attributes. Therefore, the distance-based outlier detection method to rough sets is introduced, and the advantage of rough sets is taken to solve the problem of dealing with discrete attributes. First, three distance metrics for discrete attributes within the framework of rough sets are proposed. Second, for each of these distance metrics, a corresponding outlier detection algorithm is designed, to detect outliers from data sets containing discrete attributes. Finally, the feasibility and effectiveness of these algorithms are demonstrated on two UCI data sets containing discrete attributes.

Key words: rough set theory；outlier detection；data mining；distance metric；discrete attributes

0 引 言

离群数据是数据集中偏离大部分数据的数据, 它们的表现与大多数常规对象有着明显的差异[1]. 离群数据并不等同于错误数据, 离群数据中可能蕴含着极为重要的信息, 例如在欺诈检测、网络入侵检测和灾害预测等诸多领域中, 离群点都是数据分析的主要对象[2-4].

离群点检测最早出现在统计学领域[5], 后来被引入到数据挖掘领域[2-4]. 现有的离群点检测方法大体可分为 5 类[6]: 1) 基于统计的方法[5]; 2) 基于深度的方法[7]; 3) 基于聚类的方法[8]; 4) 基于密度的方法[9]; 5) 基于距离的方法[3-4,10].

虽然人们已经提出了许多离群点检测方法, 但这些方法还存在各自的不足. 例如, 作为目前最常用的一类离群点检测方法, 基于距离的方法不能有效地处理包含离散型属性的数据集[3-4,10]. 该方法通过计算数据集中对象之间的距离来检测离群点, 离散型属性值之间并没有类似于连续型属性值之间那样固有的距离度量关系, 因此将基于距离的方法直接应用到包含离散型属性的数据集上是不合适的, 而且最终将导致检测结果存在偏差[11].

目前, 针对离散型属性的离群点检测研究还没有引起足够重视, 只有少量文献涉及到了此类问题的探讨[12-14], 而在现实生活中存在着大量的具有离散型属性的数据集, 因此研究能够处理离散型属性的离群点检测方法很必要.

自 1982 年粗糙集理论被提出以来, 现在它已经成为数据挖掘等许多领域的重要工具[15]. 但是, 目前

收稿日期: 2011-10-16；**修回日期:** 2012-03-01.

基金项目: 国家自然科学基金项目(60802042, 61103246)；山东省自然科学基金项目(ZR2011FQ005, ZR2011FQ026, ZR2010FQ027)；山东省高等学校科技计划项目(J11LG05).

作者简介: 江峰(1978−), 男, 副教授, 从事人工智能、粗糙集、数据挖掘等研究；眭跃飞(1963−), 男, 研究员, 博士生导师, 从事人工智能、数理逻辑、大规模知识处理的理论基础等研究.

粗糙集理论中关于数据挖掘问题的研究还是主要集中在分类、规则挖掘等任务上[2,16-18], 利用粗糙集进行离群点检测的研究尚不多见[19-20]. 另外, 粗糙集理论自身的特点也决定了其在处理离散型属性上具有显著的优势[15], 因此本文利用粗糙集来研究离群点的检测问题, 通过发挥粗糙集在处理离散型属性上的优势来解决传统的基于距离的离群点检测方法不能有效处理离散型属性这一问题.

1　粗糙集理论的基本知识

粗糙集理论采用一种基于信息表的知识表达形式. 信息表是一个四元组 $\mathrm{IS}=(U,A,V,f)$. 其中: U 和 A 分别表示对象集和属性集; V 是所有属性论域的并, 即 $V=\bigcup_{a\in A} V_a$, V_a 为属性 a 的值域; $f: U\times A\to V$ 是一个信息函数, 使得对于任意 $a\in A$ 以及 $x\in U, f(x,a)\in V_a$[15,18].

进一步, A 又可以划分为 2 个不相交的子集: 条件属性集 I 和决策属性集 O. 这种特殊的信息表被称为决策表, 简记为 $\mathrm{DT}=(U,I,O,V,f)$.

给定一个信息表 $\mathrm{IS}=(U,A,V,f)$, 对于任意 $B\subseteq A$, 可以定义一个论域 U 上的不可分辨关系 $\mathrm{IND}(B)$ 如下:

$$\mathrm{IND}(B)=\{(x,y)\in U\times U: \forall a\in B(f(x,a)=f(y,a))\}.$$

$\mathrm{IND}(B)$ 是 U 上的一个等价关系, 它将 U 分成多个等价类, 所有这些等价类的集合便构成了 U 的一个划分, 记为 $U/\mathrm{IND}(B)$. 对于任意 $x\in U$, 用 $[x]_B$ 表示在 $\mathrm{IND}(B)$ 下包含 x 的等价类[15,18].

2　粗糙集中的距离度量与离群点检测算法

基于距离的离群点检测方法主要通过计算对象之间的距离来检测离群点, 因此, 要想将基于距离的方法引入到粗糙集中, 利用粗糙集来解决离散型属性的处理问题, 关键在于如何在粗糙集中有效地计算任意两个对象之间的距离. 而要计算两个对象之间的距离, 就必须首先在向量空间上定义一个距离度量.

目前, 人们已经提出了很多种针对连续型属性的距离度量方法. 对于离散型属性, 由于离散型属性值之间缺少连续型属性值之间所具有的距离度量关系, 将现有的针对连续型属性的距离度量直接用于离散型属性是不合适的, 需专门为离散型属性设计相应的距离度量. 为此, 本文将定义 3 种面向离散型属性的距离度量.

2.1　粗糙集中的重叠度量与离群点检测算法

定义 1　给定信息表 $\mathrm{IS}=(U,A,V,f)$, 对于任意 $x,y\in U$, x与 y 的重叠距离 $\mathrm{OD}(x,y)$ 定义[19]如下:

$$\mathrm{OD}(x,y)=|\{a\in A: f(x,a)\neq f(y,a)\}|. \quad (1)$$

其中: $\mathrm{OD}: U\times U\to N$ 是一个从 $U\times U$ 到自然数集合 N 的函数, $|M|$ 表示集合 M 的势.

显然, 上述定义中的距离函数 OD 是一个合格的距离度量, 称之为粗糙集中的重叠度量. 下面给出一个基于重叠度量的离群点检测算法 (ODOMR).

算法 1　ODOMR.

输入: 信息表 $\mathrm{IS}=(U,A,V,f)$, 其中 $|U|=n$, $|A|=m$; 参数 p 和 d (假设距离参数 d 是一个整数, 并且 $0<d<m$).

输出: 离群点集合 L.

Step 1: 初始化: 令 $L=U$, $\mathrm{temp\_set}=\varnothing$. 并令 $P_{m-d}(A)$ 表示属性集 A 的所有具有 $m-d$ 个属性的子集所组成的一个集合, 即

$$P_{m-d}(A)=\{B\subseteq A: |B|=m-d\}.$$

Step 2: 对每个 $B\in P_{m-d}(A)$, 循环执行:

Step 2.1: 根据 B 对 U 中的所有对象进行排序;

Step 2.2: 计算划分 $U/\mathrm{IND}(B)$.

Step 3: 对 U 中的每个对象 x, 循环执行:

Step 3.1: 令 $\mathrm{temp\_set}=\varnothing$;

Step 3.2: 对每个 $B\in P_{m-d}(A)$, 循环执行:

$$\text{令 } \mathrm{temp\_set}=\mathrm{temp\_set}\bigcup[x]_B.$$

Step 3.3: 若 $|\mathrm{temp\_set}|/n>1-p$, 则令

$$L=L-\{x\}.$$

Step 4: 算法结束, 返回离群点集合 L.

算法 1 采用了一种预先对 U 中对象进行排序, 然后再计算划分 $U/\mathrm{IND}(B)$ 的方法[21], 这样可以有效降低计算划分的复杂度. 在最坏的情况下, 算法 1 的时间复杂度为 $O(C_m^d\times(m-d)\times n\log n)$, 空间复杂度为 $O(C_m^d\times n)$, 其中 m 和 n 分别为 A 和 U 的势, d 为给定的距离参数.

2.2　粗糙集中的值差异度量与离群点检测算法

前面定义的重叠度量比较直观, 但存在一个不合理的假设, 即在计算对象之间的距离时假设所有属性都具有相同的权重, 并没有考虑属性之间的差异性. 实际上, 在很多情况下这种假设是不成立的, 因为不同属性的贡献和重要性通常是不同的.

为了解决重叠度量的问题, Stanfill 等[22]提出了值差异度量 (VDM)的概念. 本文进一步将值差异度量引入粗糙集理论. 与重叠度量不同, 值差异度量基于决策表来定义.

定义 2　给定决策表 $\mathrm{DT}=(U,I,O,V,f)$, 对于任意 $x,y\in U$, x 和 y 的值差异距离 $\mathrm{VD}(x,y)$ 定义为

$$\mathrm{VD}(x,y)=\sum_{a\in I}\sqrt{D_a(x,y)}. \tag{2}$$

其中: $\mathrm{VD}: U\times U\to[0,\infty]$ 是一个从 $U\times U$ 到非负实数集的函数; $D_a(x,y)$ 表示 x 和 y 在条件属性 $a\in I$ 上的取值所产生的决策类别条件分布之间的差异, 具体定义如下:

$$D_a(x,y)=\sum_{E\in U/\mathrm{IND}(O)}\mathrm{abs}\Big(\frac{|[x]_{\{a\}}\bigcap E|}{|[x]_{\{a\}}|}-\frac{|[y]_{\{a\}}\bigcap E|}{|[y]_{\{a\}}|}\Big). \tag{3}$$

其中: 对于任意 $E\in U/\mathrm{IND}(O)$, E 中的每个对象具有相同的决策属性值, 即 E 代表了一个决策类别; 对于任意 $a\in I$, $[x]_{\{a\}}\bigcap E$ 表示等价类 $[x]_{\{a\}}$ 中所有属于决策类别 E 的对象的集合; abs 为求绝对值的函数.

定义 2 将值差异度量的思想引入到粗糙集中. 不难证明, 距离函数 VD 是一个合格的距离度量, 称之为粗糙集中的值差异度量. 该度量是对粗糙集中的重叠度量的一种有效改进, 它取消了重叠度量中所设置的不合理假设. 在计算值差异距离时, 不同的条件属性已按照对象在该属性上的取值所产生的决策类别条件分布被区别开.

下面给出一个粗糙集中基于值差异度量的离群点检测算法 (ODVDMR).

算法 2 ODVDMR.

输入: 决策表 $\mathrm{DT}=(U,I,O,V,f)$, 其中 $|U|=n$, $|I|=m$, $|O|=1$; 参数 p 和 d.

输出: 离群点集合 L.

Step 1: 初始化. 令 $L=\varnothing$, 变量 $\mathrm{count\_}x=0$, $\mathrm{count\_}a=0$, $\mathrm{count}=0$; 并令数组 $M[m,n,n]$ 中的每个元素为 0.

Step 2: 对于任意 $a\in I\bigcup O$, 循环执行:

Step 2.1: 根据属性 a, 对 U 中所有对象进行排序;

Step 2.2: 求出划分 $U/\mathrm{IND}(\{a\})$.

Step 3: 对于任意 $a\in I$, 循环执行:

Step 3.1: 对于任意 $x\in U$, 循环执行:

对于任意 $E\in U/\mathrm{IND}(O)$, 循环执行:

1) 计算 $[x]_{\{a\}}\bigcap E$;

2) 令 $M[a,x,E]=|[x]_{\{a\}}\bigcap E|$.

Step 4: 对于任意 $x\in U$, 循环执行:

Step 4.1: 令 $\mathrm{count\_}x=0$.

Step 4.2: 对于任意 $y\in U$ 且 $y\neq x$, 循环执行:

Step 4.2.1: 令 $\mathrm{count}=0$.

Step 4.2.2: 对于任意属性 $a\in I$, 循环执行:

1) 令 $\mathrm{count\_}a=0$;

2) 对于任意 $E\in U/\mathrm{IND}(O)$, 循环执行:

令 $\mathrm{count\_}a+=\mathrm{abs}\Big(\frac{|[x]_{\{a\}}\bigcap E|}{|[x]_{\{a\}}|}-\frac{|[y]_{\{a\}}\bigcap E|}{|[y]_{\{a\}}|}\Big)$;

3) 令 $\mathrm{count}+=\sqrt{\mathrm{count\_}a}$.

Step 4.2.3: 如果 $\mathrm{count}>d$, 则 $\mathrm{count\_}x++$.

Step 4.3: 如果 $\mathrm{count\_}x/n\geqslant p$, 则令 $L=L\bigcup\{x\}$.

Step 5: 算法结束, 返回离群点集合 L.

在最坏的情况下, 算法 2 的时间复杂度为 $O(m\times n^3)$, 空间复杂度为 $O(m\times n^2)$, 其中 m 和 n 分别为 I 和 U 的势.

2.3 粗糙集中的加权重叠度量与离群点检测算法

前面定义的两种距离度量各有利弊. 重叠度量比较直观, 易于理解, 但其中关于所有属性具有相同权重的假设不符合客观实际. 值差异度量取消了这种不合理假设, 但其定义过于复杂, 不易理解.

下面将上述两种距离度量的优点结合在一起, 在粗糙集中提出一种新的距离度量——加权重叠度量 (WOM). 加权重叠度量计算每个离散型属性的重要性, 并以属性的重要性作为其权值, 在计算对象的距离时, 不同的属性将按照其权重的大小发挥不同的作用.

粗糙集中关于属性重要性的计算方法有很多, 这里选择其中一种简单而且有效的方法[18].

定义 3 给定信息表 $\mathrm{IS}=(U,A,V,f)$, 对于任意 $B\subseteq A$, 令 $Q(B)=\sum_{E\in U/\mathrm{IND}(B)}|E|^2$. 对于任意 $a\in A$, 将属性 a 的重要性 $\mathrm{Sig}(a)$ 定义为

$$\mathrm{Sig}(a)=\frac{Q(A-\{a\})}{Q(A)}. \tag{4}$$

定义 4 给定信息表 $\mathrm{IS}=(U,A,V,f)$, 对于任意 $x,y\in U$, x 与 y 之间的加权重叠距离 $\mathrm{WOD}(x,y)$ 被定义为

$$\mathrm{WOD}(x,y)=\sum_{a\in A}\mathrm{Sig}(a)\times t(a,x,y). \tag{5}$$

其中: $\mathrm{WOD}: U\times U\to[0,\infty]$ 是一个从 $U\times U$ 到非负实数集的函数; $t(a,x,y): A\times U\times U\to\{0,1\}$ 是一个从 $A\times U\times U$ 到 $\{0,1\}$ 的函数, 使得对于任意 $a\in A$ 和 $x,y\in U$, 有

$$t(a,x,y)=\begin{cases}1, & f(x,a)\neq f(y,a);\\ 0, & \text{否则}.\end{cases}$$

容易证明, 上述距离函数 WOD 也是一个合格的距离度量, 称之为粗糙集中的加权重叠度量. 下面给出一个粗糙集中基于加权重叠度量的离群点检测算法 (ODWOMR).

算法 3 ODWOMR.

输入: 信息表 $\mathrm{IS}=(U,A,V,f)$, 其中 $|U|=n$, $|A|=m$; 参数 p 和 d.

输出: 离群点集合 L.

Step 1: 初始化: 令 $L=\varnothing$, 变量 $\mathrm{count}\_x=0$.

Step 2: 对于属性集 A, 执行如下操作:

Step 2.1: 根据 A, 对 U 中的所有对象进行排序;

Step 2.2: 求出划分 $U/\mathrm{IND}(A)$;

Step 2.3: 计算 $Q(A)=\sum_{E\in U/\mathrm{IND}(A)}|E|^2$.

Step 3: 对于任意 $a\in A$, 循环执行:

Step 3.1: 根据 $A-\{a\}$, 对 U 中的所有对象排序;

Step 3.2: 求出划分 $U/\mathrm{IND}(A-\{a\})$;

Step 3.3: 计算 $Q(A-\{a\})=\sum_{E\in U/\mathrm{IND}(A-\{a\})}|E|^2$;

Step 3.4: 计算 a 的重要性 $\mathrm{Sig}(a)$.

Step 4: 对于任意 $x\in U$, 循环执行:

Step 4.1: 令 $\mathrm{count}\_x=0$.

Step 4.2: 对于任意 $y\in U$ 且 $y\neq x$, 循环执行:

Step 4.2.1: 令 $\mathrm{WOD}(x,y)=0$.

Step 4.2.2: 对于任意 $a\in A$, 循环执行:

1) 令 $t(a,x,y)=0$;

2) 若 $f(x,a)\neq f(y,a)$, 则令 $t(a,x,y)=1$;

3) 令 $\mathrm{WOD}(x,y)+=\mathrm{Sig}(a)\times t(a,x,y)$.

Step 4.2.3: 若 $\mathrm{WOD}(x,y)>d$, 则 count_x++.

Step 4.3: 如果 $\mathrm{count}\_x/n\geqslant p$, 则令 $L=L\bigcup\{x\}$.

Step 5: 算法结束, 返回离群点集合 L.

在最坏的情况下, 算法 3 的时间复杂度为 $O(m\times n^2)$, 空间复杂度为 $O(m+n)$, 其中 m 和 n 分别为 A 和 U 的势.

3　实验结果

下面对前面提出的 3 种离群点检测算法的性能进行实验分析. 分别在 2 个UCI 数据集(Lymphography 和 Breast cancer) 上进行实验[23], 用以比较 ODOMR、ODVDMR、ODWOMR 算法以及传统的基于距离的算法 KNN[10]和基于神经网络的算法 RNN[24]的性能. 实验中, 对于 ODOMR 和 ODVDMR, 将它们的距离参数 d 设置为 $d=\mathrm{floor}(|A|/2)$, 其中 $\mathrm{floor}(x)$ 表示不大于 x 的最大整数; 对于 ODWOMR, 将其距离参数 d 设置为 $d=\sum_{a\in A}\mathrm{Sig}(a)\times 0.5$. 另外, 对于 KNN 算法, 设置参数 $k=5$[10].

由于在传统的基于距离的离群点检测方法中, 离群被看作一个二元属性, 缺少一个离群程度的概念[3-4,9-10]. 因此, 在实验中特别引入一个距离离群因子 (DOF)的概念, 用来刻画对象的离群程度[19-20].

3.1　Lymphography 数据集

Lymphography 数据集中包含 148 个对象和 19 个属性[23]. 所有对象被分成 4 类: normal find, metastases, malign lymph 和 fibrosis. 本文将 normal find 和 malign lymph 看作稀有类. 该数据集中共有 6 个离群点 (即属于稀有类的对象).

实验中, Lymphography 中的所有数据被导入信息表 $\mathrm{IS}_L=(U,A,V,f)$ 中, 实验结果如表 1 所示.

表 1　信息表 IS_L 中的实验结果

离群程度值前 k/% 的对象 (对象个数)	属于稀有类的对象个数 (覆盖率 /%)			
	ODOMR	ODVDMR	ODWOMR	KNN
3% (4)	4 (67)	4 (67)	4 (67)	4 (67)
3.5 %(5)	4 (67)	4 (67)	5 (83)	4 (67)
4% (6)	5 (83)	5 (83)	5 (83)	4 (67)
5.5% (8)	5 (83)	5 (83)	6 (100)	4 (67)
6% (9)	6 (100)	6 (100)	6 (100)	4 (67)
8% (12)	6 (100)	6 (100)	6 (100)	5 (83)
11% (16)	6 (100)	6 (100)	6 (100)	6 (100)

在表 1 中, 对于任意 $x\in U$, 分别利用 4 种算法计算 x 的离群程度值, 并且对于每种算法, 都会根据由该算法所计算出的 U 中对象的离群程度值, 由高到低对 U 中的对象进行排序. 因此, 在表 1 中“离群程度值前 k % 的对象 (对象个数)”是指在采用某种算法计算 U 中对象的离群程度值之后, 离群程度值排在前 k % 的对象以及这些对象的个数. 而“属于稀有类的对象个数”则是指在由该算法所检测出的离群程度值排在前 k % 的对象中, 属于稀有类的对象个数. “覆盖率”是指这些属于稀有类的对象占 U 中所有离群点的比例.

从表 1 可以看出, 对于 Lymphography 数据集, ODWOMR 算法的性能明显好于其他 3 种算法, 其中 ODOMR 与 ODVDMR 的性能相当, 而 KNN 的性能最差.

3.2　Breast cancer 数据集

Breast cancer 中包含 699 个对象和 9 个连续型属性. 所有对象被分成两类: malignant 和 benign[23]. 为了形成一个极不均匀的分布, 从该数据集中移去一些属于 malignant 类的对象[24]. 最终的数据集包括 483 个对象, 其中 39 个属于 malignant 类, 444 个属于 benign 类, 这里将 malignant 看作稀有类. 另外, 数据集中的 9 个连续型属性被分别转换成离散型属性 (最终的数据集可从以下网站获取: http: // research. cmis. csiro. au/ rohanb/outliers/breast-cancer/).

最终的 Breast cancer 数据集中的数据都被导入信息表 $\mathrm{IS}_W=(U',A',V',f')$ 中, 实验结果如表 2 所示.

从表 2 可以看出, 对于 Breast cancer 数据集, ODWOMR、ODVDMR 与 KNN 算法的性能非常接近, 并且它们的性能明显好于 ODOMR 和 RNN, RNN 的性能最差.

根据对以上实验结果的分析以及对 ODOMR、ODVDMR 和 ODWOMR 这 3 类算法的复杂性分析, 可以得出如下结论:

表 2 信息表 IS_W 中的实验结果

离群程度值前 k% 的对象(对象个数)	属于稀有类的对象个数(覆盖率/%)				
	ODOMR	ODVDMR	ODWOMR	RNN	KNN
1%(4)	4(10)	4(10)	4(10)	3(8)	4(10)
2%(8)	5(13)	7(18)	7(18)	6(15)	8(21)
4%(16)	11(28)	13(33)	14(36)	11(28)	14(36)
6%(24)	18(46)	20(51)	18(46)	18(46)	19(49)
8%(32)	24(62)	26(67)	26(67)	25(64)	25(64)
10%(40)	29(74)	33(85)	33(85)	30(77)	30(77)
12%(48)	36(92)	38(97)	37(95)	35(90)	37(95)
14%(56)	39(100)	39(100)	39(100)	36(92)	39(100)
16%(64)	39(100)	39(100)	39(100)	36(92)	39(100)
18%(72)	39(100)	39(100)	39(100)	38(97)	39(100)
20%(80)	39(100)	39(100)	39(100)	38(97)	39(100)
28%(112)	39(100)	39(100)	39(100)	39(100)	39(100)

1) ODWOMR 算法具有最好的离群点检测性能, 而且其时间复杂度比较低, 适合于处理大数据集.

2) 与 ODWOMR 和 ODVDMR 相比, ODOMR 的离群点检测性能比较差, 这是因为该算法中关于所有的属性具有相同权重的假设不合理; 与 KNN 和 RNN 相比, ODOMR 的性能在某些情况下更好一些. 另外, ODOMR 的时间复杂度最低.

3) ODVDMR 是对 ODOMR 的一种改进, 因而其离群点检测性能比 ODOMR 好. 另外, ODVDMR 的性能也比 KNN 和 RNN 好, 但 ODVDMR 的复杂度最高, 因此它不适于处理大数据集.

4 结 论

为了有效地处理离散型属性, 本文将基于距离的离群点检测方法引入粗糙集中, 提出了 3 种面向离散型属性的距离度量. 在对这 3 种距离度量的合理性和有效性进行分析的基础上, 分别设计出了 3 种基于距离的离群点检测算法. 最后, 通过实验验证了这些算法的性能. 在后续工作中将利用扩展的粗糙集模型进行离群点检测, 例如基于 Hu 等[25]提出的邻域粗糙集模型进行离群点检测.

参考文献(References)

[1] Hawkins D. Identifications of outliers[M]. London: Chapman and Hall, 1980: 1-2.

[2] Han J W, Damber M. Data mining: Concepts and technologies[M]. San Francisco: Morgan Kaufmann, 2001: 381-394.

[3] Knorr E, Ng R. Algorithms for mining distance-based outliers in large datasets[C]. Proc of the 24th VLDB Conf. New York, 1998: 392-403.

[4] Knorr E, Ng R, Tucakov V. Distance-based outliers: Algorithms and applications[J]. VLDB J: Very Large Databases, 2000, 8(3/4): 237-253.

[5] Rousseeuw P J, Leroy A M. Robust regression and outlier detection[M]. New York: John Wiley& Sons, 1987: 1-18.

[6] Kovács L, Vass D, Vidacs A. Improving quality of service parameter prediction with preliminary outlier detection and elimination[C]. Proc of the 2nd Int Workshop on Inter-Domain Performance and Simulation. Budapest, 2004: 194-199.

[7] Johnson T, Kwok I, Ng R T. Fast computation of 2-dimensional depth contours[C]. Proc of the 4th Int Conf on Knowledge Discovery and Data Mining. New York: AAAI Press, 1998: 224-228.

[8] Jain A K, Murty M N, Flynn P J. Data clustering: A review[J]. ACM Computing Surveys, 1999, 31(3): 264-323.

[9] Breunig M M, Kriegel H-P, Ng R T, et al. LOF: Identifying density-based local outliers[C]. Proc of the 2000 ACM SIGMOD Int Conf on Management of Data. Dallas: ACM Press, 2000: 93-104.

[10] Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for mining outliers from large datasets[C]. Proc of the 2000 ACM SIGMOD Int Conf on Management of Data. Dallas: ACM Press, 2000: 427-438.

[11] Chen K, Liu L. The "Best K" for entropy-based categorical data clustering[C]. Proc of the 17th Int Conf on Scientific and Statistical Database Management. 2005: 253-262.

[12] He Z Y, Deng S C, Xu X F. An optimization model for outlier detection in categorical Data[C]. Proc of Int Conf on Intelligent Computing. Hefei: Springer-Verlag, 2005: 400-409.

[13] Li S X, Lee R, Lang S D. Mining distance-based outliers from categorical data[C]. Proc of the 7th IEEE Int Conf on Data Mining Workshops. Omaha, 2007: 225-230.

[14] Narita K, Kitagawa H. Detecting outliers in categorical record databases based on attribute associations[C]. Proc of the 10th Asia-Pacific Web Conf on Progress in WWW Research and Development. Shenyang, 2008: 111-123.

[15] Pawlak Z. Rough sets[J]. Int J of Computer and Information Sciences, 1982, 11(5): 341-356.

[16] 何明, 李博, 马兆丰, 等. 粗糙集理论框架下的神经网络建模研究及应用[J]. 控制与决策, 2005, 20(7): 782-785.
(He M, Li B, Ma Z F, et al. On the neural network modeling with support rough set theory[J]. Control and Decision, 2005, 20(7): 782-785.)

[17] 代建华, 潘云鹤. 一种基于分类一致性的决策规则获取算法[J]. 控制与决策, 2004, 19(10): 1086-1090.
(Dai J H, Pan Y H. Algorithm for acquisition of decision rules based on classification consistency rate[J]. Control and Decision, 2004, 19(10): 1086-1090.)

(下转第200页)