基于粗糙集理论的序列离群点检测

江 峰[1],杜军威[1],葛 艳[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: 作为数据挖掘的一项重要任务,离群点检测已经引起人们的广泛关注.本文基于粗糙集理论来讨论离群点的定义与检测问题,提出了一种新的离群点定义——粗糙序列离群点以及相应的离群点检测算法 RSOD.该算法利用粗糙集理论中的知识熵和属性重要性等概念来构建三种类型的序列,并通过分析序列中元素的变化情况来检测离群点.在 UCI 标准数据集上,将 RSOD 算法与现有的离群点检测算法进行了比较分析,实验结果表明,我们所提出的离群点检测方法是有效的.

关键词: 离群点检测; 粗糙集; 数据挖掘; 序列; 知识熵; 属性重要性
中图分类号: TP311 **文献标识码:** A **文章编号:** 0372-2112 (2011) 02-0345-06

Sequence Outlier Detection Based on Rough Set Theory

JIANG Feng[1], DU Jun-wei[1], GE Yan[1], SUI Yue-fei[2], CAO Cun-gen[2]
(1. *College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao, Shandong* 266061, *China*;
2. *Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100080, *China*)

Abstract: As an important task of data mining, outlier detection has attracted much attention. We discuss the issues of outlier definition and detection based on rough set theory. We propose a new definition for outlier-rough sequence outlier, and the corresponding outlier detection algorithm RSOD. The algorithm constructs three kinds of sequences exploiting the notions of knowledge entropy and significance of attribute in rough sets, and detects outliers by analyzing changes of the elements in the sequences. We compare algorithm RSOD with the current outlier detection algorithms on UCI data sets. And experimental results show that our method is effective for outlier detection.

Key words: outlier detection; rough sets; data mining; sequence; knowledge entropy; significance of attribute

1 引言

作为数据挖掘的一个重要研究方向,离群点检测主要关注于数据集中的一小部分对象,与数据集中其余数据相比,这一小部分对象不符合数据集的一般模型.我们称这部分对象为离群点[1~3].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,例如在信用卡欺诈检测、网络入侵检测、疾病诊断、故障检测、灾害预测、恐怖活动防范等诸多领域中,离群点都是数据分析的主要对象[2,19,20,32,34~36].目前,对离群点的检测和分析已经发展成为数据挖掘中一项重要而又有趣的研究任务[20].

离群点检测最早出现在统计学领域[5].后来,Knorr 等将其引入到数据挖掘领域[2,3,19,32].现有的离群点检测方法主要分为 5 类:(1)基于统计的方法[5];(2)基于深度的方法[6];(3)基于聚类的方法[7];(4)基于密度的方法[8];(5)基于距离的方法[2,3,19,32,34].

本文基于粗糙集理论来研究离群点的定义与检测,提出一种基于粗糙集的序列离群点检测算法.据我们所知,利用粗糙集的方法进行离群点检测的研究还很少见[9~11,34].自 1982 年 Pawlak 提出粗糙集理论以来[17],粗糙集已经成为数据挖掘等许多领域的重要工具,但是在粗糙集理论中对于离群点检测的研究还没有引起重视.鉴于离群数据本身可能是非常重要的,离群点检测是一项有趣的数据挖掘任务.因此,本文中我们利用粗糙集来研究离群点的检测.

2 粗糙集理论的基本知识

本节简要地介绍有关粗糙集理论的一些基本概念,对于更为详细的论述可参见文献[17,31].

定义 1(近似空间) 设 U 为所讨论对象的非空有限集合,称为论域,R 为建立在 U 上的一个等价关系,

收稿日期:2009-07-30;修回日期:2010-08-06
基金项目:国家自然科学基金(No.60802042);国家 863 高技术研究发展计划(No.2007AA01Z325);山东省自然科学基金(No.ZR2009GQ013)

称二元有序组 $AS=(U,R)$ 为近似空间(Approximate Space)[17,31].

近似空间构成论域 U 的一个划分.若 R 是 U 上的一个等价关系,对任意 $x\in U$,用 $[x]_R$ 表示在等价关系 R 下包含对象 x 的等价类,U/R 表示 R 的所有等价类构成的集合,即商集.R 的所有等价类就构成 U 的一个划分.

定义 2(不分明关系) 令 $\boldsymbol{R}$ 为论域 U 上等价关系的一个族集(即由多个等价关系组成的集合),设 $\boldsymbol{P}\subseteq\boldsymbol{R}$,且 $\boldsymbol{P}\neq\emptyset$,则 $\boldsymbol{P}$ 中所有等价关系的交集称为 U 上的不分明关系(Indiscernibility Relation),记作 $IND(\boldsymbol{P})$,即有:对任意 $x\in U$,$[x]_{IND(P)}=\bigcap\limits_{R\in P}[x]_R$[17,31].

显然,$IND(\boldsymbol{P})$也是一个等价关系.这样我们就可以根据不分明关系来划分论域.

定义 3(信息表) 信息表是一个四元组 $IS=(U,A,V,f)$,其中[17,31]:

(1)U 是一个非空有限的对象集合;

(2)A 是一个非空有限的属性集合;

(3)V 是所有属性论域的并,即 $V=\bigcup\limits_{a\in A}V_a$,其中 V_a 为属性 a 的值域;

(4)$f:U\times A\to V$ 是一个信息函数,对任意 $a\in A$ 以及 $x\in U$,$f(x,a)\in V_a$.

给定一个信息表 $IS=(U,A,V,f)$,对于任意属性 $a\in A$,定义由 a 导出的二元关系如下:

$$R_a=\{(x,y)\in U\times U:f(x,a)=f(y,a)\}$$

另外,对于 A 的任意一个子集 $B\subseteq A$,定义由属性子集 B 导出的二元关系如下:

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

可以证明:对任意 $a\in A$,R_a 是一个等价关系,并且 $IND(B)=\bigcap\limits_{a\in B}R_a$,因此称 $IND(B)$为由属性集 B 导出的不分明关系[17,31].

3 基于粗糙集理论的序列离群点

基于粗糙集理论的序列离群点的基本思想是:给定一个信息表 $IS=(U,A,V,f)$,对于任意 $B\subseteq A$,如果我们针对属性集 B 采取这样一种操作,即每次从 B 中去掉当前重要性(significance)最小的那个属性,并通过观察随着关系 $IND(B)$的变化,U 中每个对象所属等价类的变化情况,就可以计算出 U 中各个对象之间的差别,从而检测出其中的离群点.因为随着 B 中元素的逐步减少,由 $IND(B)$所导致的划分 $U/IND(B)$将变得越来越粗,相应的,对象 x 的等价类 $[x]_B$ 则变得越来越大(即 $[x]_B$ 中的元素越来越多).因此,对任意 $o\in U$,如果我们逐步减少 B 中的元素,等价类 $[o]_B$ 相对于 U 中的其他对象而言,总是保持不变或变化很小,则我们认为 o 是信息表 IS 中的一个离群点.

本文的方法是对文献[11]中所提出的基于序列的离群点检测方法的有效改进.在文献[11]中,给定一个信息表 $IS=(U,A,V,f)$,我们为属性集 A 中的每个属性指派一个均匀性(variance).并且通过每次从 A 中去掉均匀性最大的那个属性,来观察 U 中每个对象所属的等价类的变化情况,进而检测出 U 中的离群点.但是,文献[11]存在这样一个问题,即为什么每次都要选择均匀性最大的那个属性从 A 中去掉,而不选择 A 中的其它属性呢?对于这一点,文献[11]并没有给出一个合理的解释.

属性重要性是粗糙集理论中的一个重要概念,在属性约简等领域有重要的应用[24~28,33].由于不同的属性在信息表 $IS=(U,A,V,f)$中所发挥的作用是不一样的,在从属性集 A 中删除属性时,我们当然希望尽量保留重要性大的属性,而优先去掉重要性小的属性,因为这样做可以尽量避免由于 A 中属性的减少所导致的信息量的损失.因此,本文在文献[11]的基础上,提出一种改进的离群点检测方法,我们为 A 中的每个属性计算其重要性,并且每次从 A 中去掉重要性最小的那个属性.最后,通过观察 U 中对象所属的等价类的变化情况来发现离群点.

为了解决信息的量化度量问题,Shannon 在 1948 年首次提出信息熵的概念[4].随着对粗糙集理论研究与应用的不断深入,粗糙集的研究人员将信息熵引入到粗糙集理论中,提出了知识熵等新的概念[14,21].

定义 4(知识熵) 给定一个信息表 $IS=(U,A,V,f)$,对于任意 $B\subseteq A$,令 $U/IND(B)=\{B_1,\cdots,B_m\}$为不分明关系(知识)$IND(B)$对论域 U 的划分.我们将知识 $IND(B)$的熵 $E(B)$定义为[14]:

$$E(B)=-\sum_{i=1}^{m}\frac{|B_i|}{|U|}\log_2\frac{|B_i|}{|U|}$$

其中,$\frac{|B_i|}{|U|}$表示论域 U 中任意对象 x 属于 B_i 的概率,$1\leqslant i\leqslant m$.$|T|$表示集合 T 的势.

定义 5(属性的重要性) 给定一个信息表 $IS=(U,A,V,f)$,对任意 $a\in A$,我们将 a 的重要性 $Sig(a)$定义为:$Sig(a)=E(A)-E(A-\{a\})$,其中 $E(A)$表示知识 $IND(A)$的熵.特别地,当 $A=\{a\}$时,$Sig(a)=E(\{a\})$.

对于任意 $a\in A$,根据定义 4 和定义 5,我们很容易证明 $0\leqslant Sig(a)\leqslant\log_2|U|$.

如果我们已经计算出 A 中每个属性的重要性,那么就可以将 A 中的所有属性按照重要性进行排序,从而得到一个属性序列(本文我们统一将序列定义成元组的形式).

定义 6(属性序列)　给定一个信息表 $IS=(U,A,V,f)$,其中 $A=\{a_1,a_2,\cdots,a_m\}$.对于任意 $a_i\in A$,我们用 $Sig(a_i)$表示 a_i 的重要性.IS 中的属性序列是一个 m -元组 $S=<a_1',a_2',\cdots,a_m'>$,其中对任意 $1\leqslant i,j\leqslant m$,$a_i'\in A$,若 $i\neq j$,则 $a_i'\neq a_j'$,并且对任意 $1\leqslant i<m$,$Sig(a_i')\leqslant Sig(a_{i+1}')$.

如果我们由属性集 A 开始,从 A 中去除重要性最小的那个属性.接下来,我们每次都从上次所获得的属性集中去除重要性最小的那个属性,直到最后得到一个只包含一个属性的集合为止.这样,就可以获得一个属性集序列[11].

定义 7(属性集序列)　给定一个信息表 $IS=(U,A,V,f)$,其中 $A=\{a_1,a_2,\cdots,a_m\}$.令 $S=<a_1',a_2',\cdots,a_m'>$ 为定义 6 中所给出的属性序列.IS 中的属性集序列是一个 m -元组 $AS=<A_1,A_2,\cdots,A_m>$,其中对任意 $1\leqslant i<m$,$A_i\subseteq A$,$A_1=A$,$A_m=\{a_m'\}$,并且对任意 $1\leqslant i<m$,$A_{i+1}=A_i-\{a_i'\}$.

根据以上定义,属性集序列 $AS=<A_1,A_2,\cdots,A_m>$ 中的任意两个相邻的属性集 A_i 和 A_{i+1}都具有这样的关系:A_{i+1}是通过将 A_i 中的属性 a_i' 去掉而得到的,其中 a_i' 是属性序列 $S=<a_1',a_2',\cdots,a_m'>$ 中的第 i 个元素,$1\leqslant i<m$.

由于属性集序列 AS 中的每个属性集都可以确定 U 上的一个不分明关系.这样,就得到 m 个 U 上的不分明关系.因此,对任意 $x\in U$,在所有 m 个不分明关系下包含 x 的等价类就形成另外一个序列——等价类序列[11].

定义 8(等价类序列)　给定一个信息表 $IS=(U,A,V,f)$,其中 $A=\{a_1,a_2,\cdots,a_m\}$.令 $AS=<A_1,A_2,\cdots,A_m>$ 为 IS 中的属性集序列,$IND(A_i)$为由 A_i 所确定的 U 上的一个不分明关系,$1\leqslant i\leqslant m$.对任意 $x\in U$,x 的等价类序列是一个 m-元组 $ES(x)=<[x]_{A_1},[x]_{A_2},\cdots,[x]_{A_m}>$,其中 $[x]_{A_i}$为在 $IND(A_i)$下包含 x 的等价类,$1\leqslant i\leqslant m$.

根据以上定义,对任意 $x\in U$,$ES(x)$中的任意两个相邻的等价类 $[x]_{A_i}$和 $[x]_{A_{i+1}}$具有这样的关系:$[x]_{A_i}\subseteq[x]_{A_{i+1}}$,$1\leqslant i<m$.

基于以上三种序列的定义,我们可以分别给出粗糙序列离群因子和粗糙序列离群点的定义.

目前的离群点检测方法普遍将离群现象看成是一种二元特性,即一个对象要么是离群点,要么不是.然而,在很多实际的场合,我们需要给每一个对象指定一个离群程度.基于这种考虑,Breunig 等引入局部离群因子的概念,用来表征每个对象的离群程度[8].类似于 Breunig 等的做法,本文提出粗糙序列离群因子的概念,用来表征对象的离群程度[9~11,34].

定义 9(粗糙序列离群因子)　给定一个信息表 $IS=(U,A,V,f)$,其中 $A=\{a_1,a_2,\cdots,a_m\}$.对于任意 $x\in U$,令 $ES(x)=<[x]_{A_1},[x]_{A_2},\cdots,[x]_{A_m}>$ 为 IS 中 x 的等价类序列.IS 中 x 的粗糙序列离群因子 $RSOF(x)$被定义为:

$$RSOF(x)=1-W(x)\times\sqrt{\frac{\sum_{j=2}^{m}\frac{\left|[x]_{A_j}\right|-\left|[x]_{A_{j-1}}\right|}{\left|[x]_{A_j}\right|}}{m-1}}$$

其中,$W:U\to[0,1)$是一个权重函数,使得对任意 $x\in U$,$W(x)=\sum_{a\in A}\frac{\left|[x]_{\{a\}}\right|}{|U|}\Big/m$ 表示对象 x 的权重.

根据以上定义,对象 x 的粗糙序列离群因子 $RSOF(x)$与其权重 $W(x)$成反比,即 x 的权重越小,则 x 越有可能成为离群点.在给定的不分明关系下,如果 x 的等价类中元素总是很少,即 U 中与 x 等价的对象总是很少,则我们认为 x 属于 U 中的一小部分对象.相应的,我们赋予 x 较小的权重,使得 x 更有可能成为离群点.因此,函数 W 体现了这样一种思想:离群点检测总是倾向于数据集中的一小部分对象,这一小部分对象比数据集中大多数对象更有可能成为离群点.

定义 10(粗糙序列离群点)　给定一个信息表 $IS=(U,A,V,f)$,其中 $A=\{a_1,a_2,\cdots,a_m\}$.令 μ 为一个给定的阈值,对任意 $x\in U$,如果 $RSOF(x)>\mu$,则 x 被称为 IS 中的一个粗糙序列离群点.

4　基于粗糙集的序列离群点检测算法 RSOD

算法 1　RSOD

输入:信息表 $IS=(U,A,V,f)$,其中 $|U|=n$,$A=\{a_1,a_2,\cdots,a_m\}$;

输出:所有粗糙序列离群点的集合 O

说明:假设 U 中的每个对象 x 都有一个标识号 id_x 用来唯一标识 x,在算法中我们将统一使用对象标识号来代表 U 中的每个对象.

(1)根据 U 中的对象在属性集 A 上的取值,按照值域 V_A 上的一个给定次序(例如字典序),对 U 中的所有对象进行排序.

(2)求出划分 $U/IND(A)=\{E_1,\cdots,E_k\}$ 中每个等价类的势,即 $|E_1|,\cdots,|E_k|$.

(3)根据 $|E_1|,\cdots,|E_k|$来计算 $IND(A)$的熵 $E(A)$.

(4)对于 A 中的每一个属性 a_i,$1\leqslant i\leqslant m$,循环执行如下操作:

(ⅰ)分别根据 U 中对象在集合 $A-\{a_i\}$ 以及属性 a_i 上的取值,按照值域 $V_{A-\{a_i\}}$与值域 V_{a_i}上的一个给定次序,对 U 中的所有对象进行排序;

(ⅱ)分别求出划分 $U/IND(A-\{a_i\})=\{F_1,\cdots,F_{s[i]}\}$ 中每个等价类的势和划分 $U/IND(\{a_i\})=\{G_1,\cdots,G_{t[i]}\}$ 中每个等价类的势;

(ⅲ)对任意 $1\leqslant j\leqslant t[i]$,循环执行

对任意 $id\_x\in G_j$,循环执行

令 $M_1[id\_x][i]=|G_j|$;

(ⅳ)根据 $|F_1|,\cdots,|F_{s[i]}|$ 来计算知识熵 $E(A-\{a_i\})$;

(ⅴ)计算属性 a_i 的重要性,即 $Sig(a_i)=E(A)-E(A-\{a_i\})$.

(5)基于属性集 A 中每个属性的重要性,构造属性序列 $S=<a_1',a_2',\cdots,a_m'>$,其中对任意 $1\leqslant i<m$,$Sig(a_i')\leqslant Sig(a_{i+1}')$.

(6)基于属性序列 S,构造属性集序列 $AS=<A_1,A_2,\cdots,A_m>$.

(7)对于属性集序列 AS 中的每一个属性集 A_i,$1\leqslant i\leqslant m$,循环执行如下操作:

(ⅰ)根据 U 中对象在属性集 A_i 上的取值,按照值域 V_{A_i} 上一个给定次序,对 U 中所有对象进行排序;

(ⅱ)求出划分 $U/IND(A_i)=\{H_1,\cdots,H_{v[i]}\}$ 中每个等价类的势;

(ⅲ)对任意 $1\leqslant j\leqslant v[i]$,循环执行

对任意 $id\_x\in H_j$,循环执行

令 $M_2[id\_x][i]=|H_j|$.

(8)对任意对象 $id\_x\in U$,循环执行如下操作:

(ⅰ)根据 $M_1[id\_x][i]$ $(1\leqslant i\leqslant m)$ 来计算权重 $W(id\_x)$;

(ⅱ)根据 $M_2[id\_x][i]$ $(1\leqslant i\leqslant m)$ 和 $W(id\_x)$,来计算离群因子 $RSOF(id\_x)$.

(ⅲ)按照离群因子的大小由高到低对 U 中所有对象进行排序.

(9)对任意对象 $id\_x\in U$,若 $RSOF(id\_x)>\mu$,则令 $O=O\cup\{id\_x\}$.

(10)算法结束,返回离群点集合 O.

在算法 1 中,我们采用了一种预先对 U 中对象进行排序,然后再计算划分 $U/IND(B)$ 的方法[18].该方法的时间复杂度为 $O(p\times n\log n)$,其中 p 和 n 分别为 B 与 U 的势,从而有效降低了计算划分 $U/IND(B)$ 的时间.

在最坏的情况下,算法 1 的时间复杂度为 $O(m^2\times n\log n)$,空间复杂度为 $O(m\times(n+m))$,其中 m 和 n 分别为集合 A 与 U 的势.

5 实验结果

本文所采用的数据集有:Lymphography 数据集和 Wisconsin Breast Cancer 数据集[16],在 Lymphography 数据集上,我们将对 RSOD 算法、SEQ 算法[11]、KNN 算法[15]和基于距离的方法[2,3]的性能进行比较,而在 Breast Cancer 数据集上,我们将比较 RSOD 算法、SEQ 算法、KNN 算法、RNN 算法[13]和基于距离的方法这五种方法的性能.关于 RNN 算法的结果可参考 Harkins 等的工作[13].

由于基于距离的方法缺少一个离群程度的概念[2,3].为了能够与 RSOD 等算法比较,实验中我们引入了一种距离离群因子的概念,具体定义参见文献[11].

5.1 Lymphography 数据集

本文将采 Aggarwal 等所提出的评价指标体系来评测每类离群点检测方法的性能,该评价体系是目前最常用的一类离群点检测方法评价体系[12].给定一个数据集以及数据集中每个对象所属的类,Aggarwal 认为要评价一个离群点检测方法的好坏,可以通过在某个给定的数据集上来运行该方法,并且计算在由该方法所找出的离群点中,真正的离群点所占据的比例.比例越高,则表明该方法的性能越好[12].

我们首先对 Lymphography 数据集进行实验[16].该数据集中包含 148 个对象,18 个符号型属性.另外,该数据集中总共有 6 个离群点(即属于稀有类的对象).实验结果如表 1 所示.

表 1 Lymphography 数据集上的结果

离群程度前 k% 的对象(对象个数)	属于稀有类的对象个数(覆盖率)			
	RSOD	SEQ	DIS	KNN
4%(6)	6(100%)	4(67%)	5(83%)	4(67%)
5%(7)	6(100%)	5(83%)	5(83%)	4(67%)
6%(9)	6(100%)	5(83%)	6(100%)	4(67%)
8%(12)	6(100%)	6(100%)	6(100%)	5(83%)
10%(15)	6(100%)	6(100%)	6(100%)	6(100%)

在表 1 中,对于 U 中的每个对象 x,我们分别利用四种算法来计算 x 的离群程度值.并且根据离群程度值由高到低对 U 中对象进行排序.因此,在表 1 中"离群程度前 k% 的对象(对象个数)"是指在采用某种算法来计算 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[9~11,30,34].

从表 1 我们可以看出,对于 Lymphography 数据集,RSOD 算法的性能明显要好于其他方法.这是因为我们在计算由各类方法所找出的离群点中真正的离群点所占据的比例时,由 RSOD 算法所得到的比例值总是最高的.例如,在各类方法所找出的离群程度值排在前 4% 的对象中(共计 6 个对象),RSOD 算法检测出的真正的

离群点数目为 6(比例达到 100%),即由 RSOD 算法所找出的 6 个离群点都是真正的离群点,没有任何误判出现,而 SEQ 算法、基于距离的方法和 KNN 算法则分别只检测出了 4、5 和 4 个真正的离群点,都存在不同程度的误判.从另外一个角度来分析,RSOD 算法只需要找出离群程度值排在前 4% 的对象就可以检测出数据集中所有的真正离群点,从而完成离群点检测的任务,而 SEQ 算法、基于距离的方法和 KNN 算法则分别需要找出离群程度值排在前 8%、6% 和 10% 的对象才可以实现这一目标.

5.2 Breast Cancer 数据集

Breast Cancer 数据集中包含 699 个对象,9 个数值型属性[16].为了形成一个非常不均匀的分布,我们仿照 Harkins 等的做法,从该数据集中移去一些属于"malignant"类的对象[13,30].最终的数据集包括 483 个对象,其中 39 个对象属于"malignant"类,444 个属于"benign"类[13].另外,我们同样将数据集中的 9 个数值型属性都转换成符号型属性*.

在最终的 Breast Cancer 数据集中,我们将"malignant"类看作稀有类.实验结果如表 2 所示.

从表 2 我们可以看出,对于 Breast Cancer 数据集,RSOD 算法、SEQ 算法和 KNN 算法的性能非常接近,它们的性能明显要好于基于距离的方法和 RNN 算法.

表 2　Breast Cancer 数据集上的结果

离群程度前 k% 的对象(个数)	属于稀有类的对象个数(覆盖率)				
	RSOD	SEQ	DIS	RNN	KNN
1% (4)	4(10%)	3(8%)	4(10%)	3(8%)	4(10%)
2% (8)	7(18%)	7(18%)	5(13%)	6(15%)	8(21%)
4% (16)	14(36%)	14(36%)	11(28%)	11(28%)	16(41%)
6% (24)	20(51%)	21(54%)	18(46%)	18(46%)	20(51%)
8% (32)	27(69%)	28(72%)	24(62%)	25(64%)	27(69%)
10% (40)	33(85%)	32(82%)	29(74%)	30(77%)	32(82%)
12% (48)	36(92%)	35(90%)	36(92%)	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%)

6　结论

本文基于粗糙集理论提出了一种新的离群点检测方法.利用粗糙集理论中的知识熵和属性重要性等概念,我们在粗糙集的信息表中给出了粗糙序列离群点的定义,并提出了相应的离群点检测算法 RSOD.通过在两个真实数据集上的实验表明,算法 RSOD 的性能要好于或等于现有的方法.由于利用粗糙集的方法进行离群点检测的研究还很少见,本文的工作扩展了粗糙集在数据挖掘等领域的应用范围,为粗糙集理论开辟了一个新的应用空间.

由于本文所提出的离群点检测方法是基于 Pawlak 的经典粗糙集模型,该模型不能直接用来处理数值型属性.为了解决这个问题,在下一步的工作中,我们计划在扩展的粗糙集模型下进行离群点检测,例如基于 Hu 等所提出的邻域粗糙集[27~29]或者模糊粗糙集[22~26]来进行离群点检测.另外,我们还计划将本文所提出的离群点检测方法作为一种无监督的入侵检测方法而应用于网络安全领域[35,36].

参考文献:

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

[2] Knorr E, Ng R. Algorithms for mining distance-based outliers in large datasets[A]. Proc of the 24th VLDB Conference[C]. New York: Morgan Kaufmann, 1998. 392 – 403.

[3] Knorr E, et al. Distance-based outliers: algorithms and applications[J]. Very Large Databases, 2000, 8(3 – 4): 237 – 253.

[4] Shannon C E. The mathematical theory of communication[J]. Bell System Technical Journal, 1948, 27(3 – 4): 373 – 423.

[5] Rousseeuw P J, Leroy A M. Robust Regression and Outlier Detection[M]. New York: John Wiley & Sons, 1987.

[6] Johnson T, et al. Fast computation of 2 – dimensional depth contours[A]. Proc of the 4th Int Conf on Knowledge Discovery and Data Mining[C]. New York: AAAI Press, 1998. 224 – 228.

[7] Jain A K, et al. Data clustering: a review[J]. ACM Computing Surveys, 1999, 31(3): 264 – 323.

[8] Breunig M M, et al. LOF: identifying density-based local outliers[A]. Proc of the 2000 ACM SIGMOD Int Conf on Management of Data[C]. Dallas: ACM Press, 2000. 93 – 104.

[9] Jiang F, et al. Outlier detection using rough set theory[A]. Proc of the 10th Int Conf on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing[C]. Canada: Springer-Verlag, 2005. 79 – 87.

[10] Jiang F, et al. A ugh set approach to outlier detection[J]. International Journal of General Systems, 2008, 37(5): 519 – 536.

[11] Jiang F, et al. Some issues about outlier detection in rough set theory[J]. Expert Systems with Applications, 2009, 36(3): 4680 – 4687.

[12] Aggarwal C C, Yu P S. Outlier detection for high dimensional data[A]. Proc of the 2001 ACM SIGMOD Int Conf on Management of Data[C]. California: ACM Press, 2001. 37 – 46.

* 最终的数据集可以从如下网站获取:
http://research.cmis.csiro.au/rohanb/outliers/breast-cancer

[13] Harkins S, et al. Outlier detection using replicator neural networks[A]. Proc of the 4th Int Conf on Data Warehousing and Knowledge Discovery[C]. France: Springer-Verlag, 2002. 170 – 180.

[14] Duntsch I, Gediga G. Uncertainty measures of rough set prediction[J]. Artificial Intelligence, 1998, 106: 109 – 137.

[15] Ramaswamy S, et al. Efficient algorithms for mining outliers from large datasets[A]. Proc of the 2000 ACM SIGMOD Int. Conf on Management of Data[C]. Dallas: ACM Press, 2000. 427 – 438.

[16] Bay S D. The UCI KDD repository[D]. Univ of California, Irvine, 1999. http://kdd.ics.uci.edu.

[17] Pawlak Z. Rough Sets, Theoretical Aspects of Reasoning about Data[M]. Dordrecht: Kluwer, 1991.

[18] Nguyen S H, Nguyen H S. Some efficient algorithms for rough set methods[A]. Proc of the 6th Int Conf on Information Processing and Management of Uncertainty[C]. Spain: Springer-Verlag, 1996. 1451 – 1456.

[19] Wang L Z, Zou L K. Research on algorithms for mining distance-based outliers[J]. Chinese Journal of Electronics, 2005, 14(3): 485 – 490.

[20] Han J W, Damber M. Data mining: concepts and technologies [M]. San Francisco: Morgan Kaufmann, 2001.

[21] Liang J Y, Shi Z Z. The information entropy, rough entropy and knowledge granulation in rough set theory[J]. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems, 2004, 12(1): 37 – 46.

[22] Jensen R, Shen Q. Semantics-Preserving dimensionality reduction: Rough and fuzzy-rough-based approaches [J]. IEEE Trans on Knowledge and Data Engineering, 2004, 16(12): 1457 – 1471.

[23] Dubois D, Prade H. Rough fuzzy sets and fuzzy rough sets [J]. International Journal of General Systems, 1990, 17: 191 – 209.

[24] Hu Q H, et al. Fuzzy probabilistic approximation spaces and their information measures[J]. IEEE Trans on Fuzzy Systems, 2006, 14(2): 191 – 201.

[25] Hu Q H, et al. Information-Preserving hybrid data reduction based on fuzzy-rough techniques[J]. Pattern Recognition Letters, 2006, 27(5): 414 – 423.

[26] Hu Q H, et al. Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation [J]. Pattern Recognition, 2007, 40(12): 3509 – 3521.

[27] Hu Q H, et al. Neighborhood classifiers[J]. Expert Systems with Applications, 2008, 34(2): 866 – 876.

[28] Hu Q H, et al. Mixed feature selection based on granulation and approximation[J]. Knowledge-Based Systems, 2008, 21(4): 294 – 304.

[29] Hu Q H, et al. Neighborhood rough set based heterogeneous feature subset selection[J]. Information Sciences, 2008, 178(18): 3577 – 3594.

[30] He Z Y, et al. A fast greedy algorithm for outlier mining[A]. Proceedings of PAKDD 2006[C]. Singapore: Springer-Verlag, 2006. 567 – 576.

[31] 刘清. Rough 集及 Rough 推理[M]. 北京: 科学出版社, 2001.

[32] 黄毅群, 卢正鼎, 等. 分布式异常检测中隐私保持问题研究[J]. 电子学报, 2006, 34(5): 796 – 799.
Huang Y Q, Lu Z D, et al. Privacy preserving outlier detection [J]. Acta Electronica Sinica, 2006, 34(5): 796 – 799. (in Chinese)

[33] 邓大勇, 黄厚宽, 等. 不一致决策系统中约简之间的比较[J]. 电子学报, 2007, 35 (2): 252 – 255.
Deng D Y, Huang H K, et al. Comparison of various types of reductions in inconsistent systems[J]. Acta Electronica Sinica, 2007, 35(2): 252 – 255. (in Chinese)

[34] 江峰, 杜军威, 等. 基于边界和距离的离群点检测[J]. 电子学报, 2010, 38(3): 700 – 705.
Jiang F, Du J W, et al. Outlier detection based on boundary and distance[J]. Acta Electronica Sinica, 2010, 38(3): 700 – 705. (in Chinese)

[35] 陶新民, 陈万海, 等. 一种新的基于模糊聚类和免疫原理的入侵检测模型[J]. 电子学报, 2006, 34(7): 1329 – 1332.
Tao X M, ChenWH, et al. A novel model of IDS based on fuzzy cluster and immune principle[J]. Acta Electronica Sinica, 2006, 34(7): 1329 – 1332. (in Chinese)

[36] 罗敏, 王丽娜, 等. 基于无监督聚类的入侵检测方法[J]. 电子学报, 2003, 31(11): 1713 – 1716.
Luo M, Wang L N, et al. An unsupervised clustering-based intrusion detection method[J]. Acta Electronica Sinica, 2003, 31(11): 1713 – 1716. (in Chinese)

作者简介:

江 峰 男, 1978 年生, 博士、副教授. 主要研究方向为粗糙集理论、人工智能、网络安全. E-mail: jiangkong@163.net

杜军威 男, 1974 年生, 博士、副教授. 主要研究方向为人工智能、软件的可信性分析与验证.

葛 艳 女, 1975 年生, 博士、副教授. 主要研究方向为人工智能、智能控制、模糊集理论等.

眭跃飞 男, 1963 年生, 研究员, 博士生导师, 中国计算机学会高级会员. 主要研究方向为人工智能、数理逻辑、大规模知识处理的理论基础.