基于边界和距离的离群点检测

江 峰[1],杜军威[1],眭跃飞[2],曹存根[2]
(1.青岛科技大学信息与科学技术学院,山东青岛 266061; 2.中国科学院计算技术研究所,北京 100080)

摘 要: 近年来,离群点检测已经引起人们的广泛关注.离群点检测在网络入侵检测、信用卡欺诈、电子商务犯罪、医疗诊断以及反恐等诸多领域都具有十分重要的作用.离群点检测的目的是为了发现数据集中的一小部分对象,与数据集中其余的大部分对象相比,这一小部分对象有着特殊的行为或者具有反常的属性.针对现有的离群点检测方法不能有效处理不确定与不完整数据的问题,本文将粗糙集中边界的概念与 Knorr 等所提出的基于距离的离群点检测方法结合在一起,在粗糙集的框架中提出一种新的离群点定义与检测方法.针对于该方法,我们设计出相应的离群点检测算法 BDOD,并且通过在临床诊断数据集上所进行的实验,验证了算法 BDOD 的有效性.实验结果表明本文的方法为处理离群点检测中的不确定与不完整数据问题提供了一条新的途径.

关键词: 数据挖掘; 离群点检测; 粗糙集; 不确定与不完整数据
中图分类号: TP274 **文献标识码:** A **文章编号:** 0372-2112 (2010) 03-0700-06

Outlier Detection Based on Boundary and Distance

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

Abstract: In recent years, outlier detection has gained considerable interest. The identification of outliers is important for many applications such as intrusion detection, credit card fraud, criminal activities in electronic commerce, medical diagnosis and anti-terrorism, etc. The aim of outlier detection is to find small groups of objects who behave in an unexpected way or have abnormal properties when compared with the rest large amount of data. Since the existing methods for outlier detection cannot deal with uncertain and incomplete data. In this paper, we propose a new method for outlier definition and detection, which exploits the basic notion —boundary of rough sets and Knorr's method about distance-based outliers. We also give an algorithm BDOD to find such outliers within the framework of rough set theory. The effectiveness of our algorithm is demonstrated on publicly clinical diagnosis data sets. Our method gives a new approach to the treatment of uncertain and incomplete data in outlier detection.

Key words: data mining; outlier detection; rough sets; uncertain and incomplete data

1 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,以至于让人怀疑它们可能是由另外一种完全不同的机制所产生的[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,例如在信用卡欺诈检测、网络入侵检测、疾病诊断、通信欺诈分析、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[2,3].在所有的科学研究领域中,离群数据都可能给予我们新的视角,从而导致新的理论和新的应用的不断出现.因此,对离群数据进行分析与研究具有十分重要的理论意义和实际应用价值.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[3].

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[2,18,19,21].现有的离群点检测方法主要有五类[4]:(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[2,18,19,21].经过分析,我们发现这些方法基本上都是采用确定性的方式来表示和处理数据的,并没有考虑数据的不确定与不完整性问题.而我们的现实生活中又存在着大量不确定与不完整数据.对于这种类型的数据,现有的离群点检测方法还无法处理.因此,我们迫切需要一种能够处理不确定与不完整数据的离群点检测方法.

收稿日期:2008-12-22;修回日期:2009-03-23
基金项目:国家自然科学基金(No.60802042,60674004,60641010,60573063,60573064);国家863高技术研究发展计划(No.2007AA01Z325);青岛科技大学引进人才启动基金(No.200702583)

针对上述问题,在前期研究工作中,本文作者深入研究了如何利用粗糙集来进行离群点检测的问题,并提出了若干基于粗糙集的离群点检测方法[9~11].在文献[9]中,基于粗糙集边界的概念,我们提出了一种基于边界的离群点检测方法.另外,在论文[11]中,我们将基于距离的离群点检测方法引入到粗糙集中,并提出了两种针对分类型属性的距离度量,用于计算对象之间的距离.

本文将在前期工作基础上,进一步把基于边界的与基于距离的离群点检测方法结合在一起,在粗糙集的框架中提出一种基于边界和距离的离群点检测方法.自 1982 年 Pawlak 提出粗糙集理论以来[16],粗糙集作为处理不确定与不完整数据的重要工具,受到广泛关注.经过二十余年的发展,粗糙集已成为数据挖掘、机器学习等领域的重要方法,其中数据约简是其最主要的贡献之一[22].但是,目前在粗糙集理论中对于离群点检测的研究还没有引起足够的重视,类似的研究还很少见.因此,本文利用粗糙集理论来研究离群点检测,选题具有较强的创新性.由于我们的现实世界中存在着大量不确定与不完整数据,离群点检测不可避免地会遇到不确定与不完整数据的处理问题,因此,本文的研究不仅可以为离群点检测中的不确定与不完整数据的处理提供一种新的解决办法,而且还可以拓宽粗糙集理论在数据挖掘等领域的应用范围,为粗糙集理论开辟一个新的应用空间.

2　粗糙集理论的基本知识

粗糙集理论采用基于信息表的知识表示形式,信息表是粗糙集理论表示和处理知识的基本工具.信息表通常被定义成一个四元组 $IS=(U,A,V,f)$,其中 U 和 A 分别代表对象集合与属性集合;V 是所有属性论域的并集;f 是一个信息函数,使得对任意 $a\in A$ 和 $x\in U$,$f(x,a)\in V$[16].

给定一个信息表 $IS=(U,A,V,f)$,对任意的属性子集 $B\subseteq A$,我们都可以确定论域 U 上的一个不可区分关系 $IND(B)=\{(x,y)\in U\times U:\forall a\in B(f(x,a)=f(y,a))\}$[16].关系 $IND(B)$ 将论域 U 划分成多个等价类,所有这些等价类就构成 U 的一个划分,记为 $U/IND(B)$.对任意对象 $x\in U$,本文将使用 $[x]_B$ 来表示在关系 $IND(B)$ 下包含对象 x 的等价类[16,20].

定义 1　给定一个信息表 $IS=(U,A,V,f)$,对于任意 $B\subseteq A$ 和 $X\subseteq U$,X 的 B-上近似和 B-下近似分别被定义为:

$$\overline{X}_B=\bigcup\{[x]_B\in U/IND(B):[x]_B\cap X\neq\emptyset\};$$

$$\underline{X}_B=\bigcup\{[x]_B\in U/IND(B):[x]_B\subseteq X\}.$$

另外,$BN_B(X)=\overline{X}_B-\underline{X}_B$ 被称为集合 X 的 B-边界.我们可以将 X 的边界看成是在现有的知识条件下,无法对其进行确定分类的那些元素所组成的集合.边界是某种意义上论域 U 中的不确定域.因此,相对于 U 中的其它对象而言,边界中的元素是一类特殊的对象,这些元素既不能确定地属于 X,也不能确定地不属于 X[16,20].

既然相对于 U 中其它对象而言,边界中的元素是一类特殊的对象,而我们在进行离群点检测时,正好需要在给定数据集中寻找一小部分行为比较特殊或者具有反常属性的对象.因此,本文在讨论离群点检测时,将考虑使用集合边界所蕴含的信息来进行离群点检测[9].

3　基于边界和距离的离群点

本文将针对信息表来设计基于边界和距离的离群点检测方法,该方法的主要思想可以描述如下:

给定一个信息表 $IS=(U,A,V,f)$ 和任意 $X\subseteq U(X\neq\emptyset)$.对于任意 $B\subseteq A$,首先,根据关系 $IND(B)$ 将集合 X 分成三个部分:异常边界 $EB(X)$、B-主边界 $PB_B(X)$ 和 B-下近似 $\underline{X}_B$.然后,针对任意 $x\in X$,分别计算 x 与 $EB(X)$、$PB_B(X)$ 以及 $\underline{X}_B$ 中每个对象之间的距离.最后,根据所求得的距离值,就可以判断 x 是否是一个离群点.

虽然上述方法也是通过计算对象 x 与 X 中所有对象的距离来判定 x 是否为离群点.但是,与基于距离的离群点检测不同的是[2,18,19],我们在寻找 X 中的离群点时,首先将 X 分成三个部分,然后对来自这三个不同部分的对象采取不同的方式进行处理.具体来说,对于异常边界中的对象,我们认为这些对象是离群点的可能性最大.因此,如果异常边界中存在越多的对象与 x 的距离较近,则 x 越有可能是离群点.而对于下近似中的对象,我们认为这些对象是离群点的可能性最小.因此,如果下近似中存在越多的对象与 x 的距离较远,则 x 越有可能是离群点.另外,对于主边界中的对象,我们认为这些对象是离群点的可能性居中.因此,如果主边界中存在越多的对象与 x 保持适当的距离,则 x 越有可能是离群点.总之,在给定的知识条件下,如果对象 x 总是与异常边界中的对象靠得很近,而与下近似中的对象离得很远,并且与主边界中的对象保持适当的距离,则我们认为 x 是 X 中的一个基于边界和距离的离群点.

在传统的基于距离的离群点检测方法中,给定数据集 X 和 $x\in X$,只要 X 中的大部分(超过一定比例)的对象与 x 的距离较远(大于给定的阈值),就认为 x 是一个离群点[2,18,19].虽然这种方法比较简单,但它却忽略了 X 中对象之间的差异.如果我们在检测离群点时,采用同一种方式来处理 X 中的所有对象,不加以区分,

明显这是不合理的,并且最终将导致检测结果存在着偏差.因此,本文所提出的基于边界和距离的离群点检测方法是对传统的基于距离方法的一种改进.

定义 2(内边界) 给定一个信息表 $IS=(U,A,V,f)$和任意的 $X\subseteq U(X\neq\emptyset)$.对于任意 $B\subseteq A$,我们将集合 X 的 B-内边界定义为:

$$IB_B(X)=\bigcup\{x\in X:[x]_B\not\subseteq X\}$$

命题 1 给定一个信息表 $IS=(U,A,V,f)$和任意的 $X\subseteq U(X\neq\emptyset)$.对于任意 $B\subseteq A$,令 $IB_B(X)$和 $\underline{X}_B$ 分别为 X 的 B-内边界和 B-下近似,则 $IB_B(X)=X-\underline{X}_B$.

证明 由于 $\underline{X}_B=\bigcup\{x\in X:[x]_B\subseteq X\}$, $IB_B(X)=\bigcup\{x\in X:[x]_B\not\subseteq X\}$,并且对于任意 $x\in X$,$[x]_B\subseteq X$ 或者 $[x]_B\not\subseteq X$.因此,$x\in\underline{X}_B$ 或者 $x\in IB_B(X)$,即 $x\in IB_B(X)\cup\underline{X}_B$,所以 $X\subseteq IB_B(X)\cup\underline{X}_B$.另外,由内边界和下近似的定义可知,$\underline{X}_B\subseteq X$ 且 $IB_B(X)\subseteq X$,因此 $IB_B(X)\cup\underline{X}_B\subseteq X$.这样,我们就有得到 $IB_B(X)\cup\underline{X}_B=X$.

另外,不存在一个对象 $x\in X$,使得 $[x]_B\subseteq X$ 且 $[x]_B\not\subseteq X$,即不存在一个对象 $x\in X$ 使得 $x\in\underline{X}_B$ 且 $x\in IB_B(X)$.因此,$IB_B(X)\cap\underline{X}_B=\emptyset$.

由 $IB_B(X)\cup\underline{X}_B=X$ 和 $IB_B(X)\cap\underline{X}_B=\emptyset$,我们可以得到 $IB_B(X)=X-\underline{X}_B$.

根据上述命题,对于任意的 $X\subseteq U$ 和 $B\subseteq A$,我们都可以把 X 分成两个部分:B-内边界和 B-下近似.此外,我们还可以进一步把 X 的 B-内边界分成两个部分:异常边界和主边界.

定义 3(异常边界) 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 $X\subseteq U(X\neq\emptyset)$ 和任意 $a_i\in A$,令 $IB_{\{a_i\}}(X)$为 X 的$\{a_i\}$-内边界,$1\leqslant i\leqslant m$.我们将集合 X 在信息表 IS 中的异常边界定义为:

$$EB(X)=\bigcap_{i=1}^{m}IB_{\{a_i\}}(X)$$

定义 4(主边界) 给定一个信息表 $IS=(U,A,V,f)$和任意的 $X\subseteq U(X\neq\emptyset)$.对于任意 $B\subseteq A$,令 $IB_B(X)$和 $EB(X)$分别为 X 的 B-内边界和异常边界.我们将集合 X 的 B-主边界定义为:

$$PB_B(X)=IB_B(X)-EB(X)$$

定义 5(偏离因子) 给定一个信息表 $IS=(U,A,V,f)$和任意的 $X\subseteq U(X\neq\emptyset)$.对于任意 $B\subseteq A$ 和 $x\in X$,我们将对象 x 相对于集合 X 的 B-偏离因子定义为:

$$DF_X^B(x)=\left(\begin{array}{l}|\{y\in EB(X):d(x,y)\leqslant d_1\}|+\\|\{y\in PB_B(X):d(x,y)\geqslant d_2\}|+\\|\{y\in\underline{X}_B:d(x,y)\geqslant d_3\}|\end{array}\right)\Big/|X|$$

其中 $d(x,y)$为在某个给定的距离度量下对象 x 与 y 间的距离[2,3].另外,d_1、d_2 和 d_3 是三个给定的距离阈值.

对象 x 的偏离因子 $DF_X^B(x)$体现了 x 在现有知识条件下,是一个离群点的可能性.为了刻画数据集中每个对象的离群程度,本文将在偏离因子的基础上引入一个多重离群因子(Multiple Outlier Factor,MOF)的概念,用来表征信息表中每个对象的离群程度[8,10,11].

定义 6(多重离群因子) 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 $X\subseteq U(X\neq\emptyset)$和任意 $x\in X$,我们将对象 x 相对于集合 X 的多重离群因子 $MOF_X(x)$ 定义为:

$$MOF_X(x)=\frac{\sum_{j=1}^{m}DF_X^{\{a_j\}}(x)\times W_X^{\{a_j\}}(x)}{|A|}$$

其中,$DF_X^{\{a_j\}}(x)$为对象 x 相对于 X 的$\{a_j\}$-偏离因子;$W_X^{\{a_j\}}:X\to[0,1)$是一个权重函数,使得对任意 $x\in X$,$W_X^{\{a_j\}}(x)=1-\sqrt{\frac{|[x]_{a_j}\cap X|}{|X|}}$ 为 x 的权重,$1\leqslant j\leqslant m$.$|M|$表示集合 M 的势.

定义 7(基于边界和距离的离群点) 给定一个信息表 $IS=(U,A,V,f)$和任意的 $X\subseteq U(X\neq\emptyset)$.令 μ 为一个给定的阈值,对于任意 $x\in X$,如果 $MOF_X(x)>\mu$,则 x 被称为 X 中的一个基于边界和距离的离群点,其中 $MOF_X(x)$为对象 x 相对于集合 X 的多重离群因子.

4 基于边界和距离的离群点检测算法 BDOD

算法 1 BDOD

输入 信息表 $IS=(U,A,V,f)$和 $X\subseteq U$,其中$|U|=n$,$A=\{a_1,a_2,\cdots,a_m\}$,$|X|=n_X$.阈值 μ、d_1、d_2 和 d_3

输出 X 中所有离群点的集合 O

(1)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) 根据 U 中对象在属性 a_i 上的取值,按照值域上的一个给定次序(例如字典序),对 U 中的所有对象进行排序[17];

(ⅱ)求出划分 $U/IND(\{a_i\})$;

(ⅲ)计算 X 的$\{a_i\}$-内边界和$\{a_i\}$-下近似.

(2)计算 X 的异常边界.

(3)对于任意 $1\leqslant i\leqslant m$,计算 X 的$\{a_i\}$-主边界.

(4)对于 X 中的每个对象 x,循环执行如下操作:

(ⅰ)对于任意 $y\in X$,计算对象 x 与 y 之间的距离 $d(x,y)$;

(ⅱ)对于任意 $1\leqslant i\leqslant m$,计算 x 相对于 X 的$\{a_i\}$-偏离因子和$\{a_i\}$-权重;

(ⅲ)计算对象 x 相对于 X 的多重离群因子 $MOF_X(x)$;

(ⅳ)如果 $MOF_X(x)>\mu$,则令 $O=O\cup\{x\}$.

(5)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[17],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 $O((m\times n_X^2)+(m\times n\log n))$,空间复杂度为 $O(m\times n)$,其中 m,n 和 n_X 分别为集合 A,U 与 X 的势.

5　实验结果

为了验证 BDOD 算法的有效性,我们将通过实验来比较 BDOD 算法、基于边界的离群点检测方法[9]和基于距离的离群点检测方法[11]各自的性能.在实验中,对于 BDOD 算法,我们将采用"基于粗糙集的覆盖度量"作为距离度量[11].另外,我们将 d_1、d_2 和 d_3 这三个距离阈值分别设置为:$d_1=|A|/3$,$d_2=|A|/2$,$d_3=0.9\times|A|$,其中 $|A|$ 代表属性集 A 的势.对于基于边界的离群点检测方法和基于距离的离群点检测方法,具体的实验细节请参考文献[10].

实验中所采用的数据集有 2 个:Lymphography(淋巴系统造影术)数据集和 Wisconsin Breast Cancer(威斯康星乳腺癌)数据集[15].在这两个数据集上,我们将采用 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[12,14].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].

5.1　Lymphography 数据集

Lymphography 数据集中包含 148 个对象和 19 个属性[15].所有的对象被分成四个类:"normal find"、"metastases"、"malign lymph"和"fibrosis".我们将"normal find"和"malign lymph"看作稀有类(注:属于稀有类的对象都是离群点).

在实验中,Lymphography 数据集中的所有数据都被导入到信息表 $IS_L=(U,A,V,f)$ 中.我们分别在 U 的两个子集 X_1 和 X_2 中检测离群点,其中:(1) $X_1=\{x\in U: f(x,dislocation)=1\}$;(2) $X_2=\{x\in U: f(x,early\_uptake)=1\vee f(x,bl\_affere)=1\}$.具体的实验结果如下面的表 1 所示.

表 1　信息表 IS_L 中关于 X_1 和 X_2 的实验结果

$X_1:\|X_1\|=50,\|R_{X_1}\|=4$				$X_2:\|X_2\|=90,\|R_{X_2}\|=5$			
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)			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)		
	BDOD	DIS	BOU		BDOD	DIS	BOU
2%(1)	1(25%)	1(25%)	1(25%)	2%(2)	2(40%)	2(40%)	2(40%)
4%(2)	2(50%)	2(50%)	2(50%)	4%(4)	4(80%)	3(60%)	3(60%)
6%(3)	3(75%)	3(75%)	2(50%)	5%(5)	4(80%)	4(80%)	3(60%)
8%(4)	4(100%)	3(75%)	2(50%)	8%(7)	5(100%)	4(80%)	3(60%)
10%(5)	4(100%)	3(75%)	2(50%)	14%(13)	5(100%)	5(100%)	3(60%)
12%(6)	4(100%)	4(100%)	2(50%)	66%(59)	5(100%)	5(100%)	4(80%)
32%(16)	4(100%)	4(100%)	3(75%)	70%(63)	5(100%)	5(100%)	5(100%)
40%(20)	4(100%)	4(100%)	4(100%)				

在表 1 中,"BDOD"、"DIS"和"BOU"分别代表 BDOD 算法、基于距离的和基于边界的离群点检测方法.$|X_j|$ 和 $|R_{X_j}|$ 分别表示集合 X_j 中的元素个数以及 X_j 中的离群点个数,$1\leqslant j\leqslant 2$.对于 X_j 中的每个对象 x,我们分别利用这三种离群点检测方法来计算 x 的离群程度值.然后根据每种方法所计算出的 X_j 中对象的离群程度值,由高到低对 X_j 中对象进行排序.因此,在表 1 中"离群程度值前 $k\%$ 的对象(对象个数)"是指在采用某种离群点检测方法来计算 X_j 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"则是指在由该方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 X_j 中所有离群点的比例,$1\leqslant j\leqslant 2$[10,11,14].

从表 1 中我们可以看出,对于 Lymphography 数据集,BDOD 算法的性能明显要好于基于距离的方法和基于边界的方法,其中基于边界的方法的性能最差.

5.2　Breast Cancer 数据集

Breast Cancer 数据集中包含 699 个对象和 9 个连续型属性.所有对象被分成两类:"malignant"和"benign"[15].为了形成一个极不均匀的分布,我们从该数据集中移去一些属于"malignant"类的对象[13].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.另外,数据集中的 9 个连续型属性被分别转换成分类型属性*[13-14].

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

在最终所获得的 Breast Cancer 数据集中,我们将"malignant"类看作稀有类.另外,我们将数据集中的数据都导入到信息表 $IS_W = (U', A', V', f')$ 中[10,11].我们分别在 U' 的两个子集 X_1' 和 X_2' 中检测离群点,其中:(1) $X_1' = \{x \in U' : f'(x, \text{Clump\_thickness}) = 5\}$;(2) $X_2' = \{x \in U' : f'(x, \text{Mitoses}) = 1\}$.具体的实验结果如表 2 所示.从表 2 中我们可以看出,对于 Breast Cancer 数据集中,BDOD 算法的性能也明显要好于基于距离的方法和基于边界的方法.因此,这同样证明了我们的方法的有效性.

表 2 信息表 IS_W 中关于 X_1' 和 X_2' 的实验结果

X_1': $\lvert X_1' \rvert = 87$, $\lvert R_{X_1'} \rvert = 4$				X_2': $\lvert X_2' \rvert = 454$, $\lvert R_{X_2'} \rvert = 23$			
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)			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)		
	BDOD	DIS	BOU		BDOD	DIS	BOU
2%(2)	2(50%)	2(50%)	2(50%)	1%(5)	4(17%)	4(17%)	4(17%)
3%(3)	3(75%)	2(50%)	3(75%)	2%(9)	8(35%)	6(26%)	7(30%)
5%(4)	3(75%)	3(75%)	3(75%)	3%(14)	11(48%)	10(43%)	11(48%)
6%(5)	4(100%)	3(75%)	3(75%)	4%(18)	14(61%)	12(52%)	13(56%)
7%(6)	4(100%)	4(100%)	3(75%)	5%(23)	18(78%)	15(65%)	18(78%)
8%(7)	4(100%)	4(100%)	4(100%)	6%(27)	20(87%)	18(78%)	20(87%)
				7%(32)	23(100%)	23(100%)	21(91%)
				10%(45)	23(100%)	23(100%)	22(96%)
				12%(54)	23(100%)	23(100%)	23(100%)

6 结论

针对当前的离群点检测方法无法处理不确定与不完整数据的问题,本文将基于粗糙集边界的离群点检测方法与传统的基于距离的离群点检测方法结合在一起,充分发挥这两类方法各自的特点,提出了一种基于边界和距离的离群点检测方法.该方法利用粗糙集在处理不确定与不完整数据方面的优势,可以从不确定与不完整的数据中高效地检测出离群点.针对该方法,我们在粗糙集的信息表中设计出相应的离群点检测算法 BDOD,并且通过实验表明,基于边界和距离的方法比基于边界的方法以及基于距离的方法具有更好的性能.

由于利用粗糙集的方法进行离群点检测的研究还很少见,本文的工作不仅使得离群点检测可以处理不确定与不完整的数据,而且还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.在下一步的工作中,我们打算将本文所提出的离群点检测方法应用于网络入侵检测,用来解决现有的入侵检测系统中所普遍存在的检测准确率低、误警率高的问题[23].

参考文献:

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

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

[3] J W Han, M Damber. Data Mining: Concepts and Technologies [M]. San Francisco: Morgan Kaufmann, 2001.

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

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

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

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

[8] M M Breunig, H-P Kriegel, R T Ng, J Sander. LOF: identifying density-based local outliers[A]. In Proc of the 2000 ACM SIGMOD Int Conf on Management of Data[C]. Dallas: ACM Press, 2000. 93 – 104.

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

[10] F Jiang, Y F Sui, C G Cao. A rough set approach to outlier detection[J]. International Journal of General Systems, 2008, 37(5): 519 – 536.

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

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

46.
[13] S Harkins, HXHe, G J Williams, R A Baxter. Outlier detection using replicator neural networks[A]. In Proc of the 4th Int Conf on Data Warehousing and Knowledge Discovery[C]. France: Springer-Verlag, 2002. 170 – 180.
[14] Z Y He, S C Deng, XF Xu. An optimization model for outlier detection in categorical data[A]. In Int Conf on Intelligent Computing[C]. China: Springer-Verlag, 2005. 400 – 409.
[15] S D Bay. The UCI KDD repository[DB]. http://kdd.ics.uci.edu, 1999.
[16] Z Pawlak, Rough Sets. Theoretical Aspects of Reasoning about Data[M]. Dordrecht: Kluwer, 1991.
[17] S H Nguyen, H S Nguyen. Some efficient algorithms for rough set methods[A]. In Proc of the 6th Int Conf on Information Processing and Management of Uncertainty [C]. Spain: Springer-Verlag, 1996. 1451 – 1456.
[18] L Z Wang, L K Zou. Research on algorithms for mining distance-based outliers[J]. Chinese Journal of Electronics, Beijing, 14(3), 2005. 485 – 490.
[19] E Knorr, R Ng, V Tucakov. Distance-based outliers: algorithms and applications[J]. VLDB Journal, 2000, 8(3 – 4): 237 – 253.
[20] 刘清. Rough 集及 Rough 推理[M]. 北京:科学出版社, 2001.
Q Liu. Rough Sets and Rough Reasoning[M]. Beijing: Science Press, 2001. (in Chinese)
[21] 黄毅群,卢正鼎,胡和平,李瑞轩. 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.
Y Q Huang, Z D Lu, H P Hu, RXLi. Privacy preserving outlier detection[J]. Acta Electronica Sinica, 2006, 34(5): 796 – 799. (in Chinese)
[22] 邓大勇,黄厚宽,李向军. 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.
D Y Deng, H K Huang, X J Li. Comparison of various types of reductions in inconsistent systems[J]. Acta Electronica Sinica, 2007, 35(2): 252 – 255. (in Chinese)
[23] 陶新民,陈万海,郭黎利. 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.
X M Tao, W H Chen, L L Guo. A novel model of IDS based on fuzzy cluster and immune principle[J]. Acta Electronica Sinica 2006, 34(7): 1329 – 1332. (in Chinese)

作者简介:

江　峰　男,1978 年生,博士、副教授. 2007 年毕业于中科院计算所. 主要研究方向有粗糙集理论、人工智能. 现主持国家自然科学基金项目 1 项. 近年来,发表论文 10 多篇,其中 SCI 收录 6 篇.
E-mail: jiangkong@163.net

眭跃飞　男,1963 年生,中科院计算所研究员,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.

曹存根　男,1964 年出生,中科院计算所研究员,博士生导师,入选中科院百人计划. 主要研究方向为人工智能、知识工程、大规模知识获取与知识处理、情感计算等.