控制与决策
Control and Decision

基于粗糙熵的离群点检测方法及其在无监督入侵检测中的应用

江峰， 王凯郦， 于旭， 眭跃飞， 杜军威

引用本文:
江峰, 王凯郦, 于旭, 等. 基于粗糙熵的离群点检测方法及其在无监督入侵检测中的应用[J]. 控制与决策, 2020, 35(5): 1199–1204.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2018.1345

文章编号: 1001-0920(2020)05-1199-06 DOI: 10.13195/j.kzyjc.2018.1345

基于粗糙熵的离群点检测方法及其在无监督入侵检测中的应用

江 峰[1], 王凯郦[1], 于 旭[1], 眭跃飞[2], 杜军威[1†]

(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: 香农的信息熵被广泛用于粗糙集. 利用粗糙集中的粗糙熵来检测离群点,提出一种基于粗糙熵的离群点检测方法,并应用于无监督入侵检测. 首先,基于粗糙熵提出一种新的离群点定义,并设计出相应的离群点检测算法——基于粗糙熵的离群点检测(rough entropy-based outlier detection, REOD);其次,通过将入侵行为看作是离群点,将REOD应用于入侵检测中,从而得到一种新的无监督入侵检测方法. 通过多个数据集上的实验表明,REOD具有良好的离群点检测性能. 另外,相对于现有的入侵检测方法,REOD具有较高的入侵检测率和较低的误报率,特别是其计算开销较小,适合于在海量高维的数据中检测入侵.

关键词: 离群点检测；粗糙集；粗糙度；粗糙熵；无监督入侵检测

中图分类号: TP391 **文献标志码:** A

A rough entropy-based approach to outlier detection and its application in unsupervised intrusion detection

JIANG Feng[1], *WANG Kai-li*[1], *YU Xu*[1], *SUI Yue-fei*[2], *DU Jun-wei*[1†]

(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)

Abstract: The information entropy, proposed by Shannon, has been widely used in rough sets. In this paper, we use the rough entropy in rough sets to detect outliers, and propose a rough entropy-based outlier detection approach, which is applied to unsupervised intrusion detection. Firstly, we propose a new definition for outliers based on rough entropy, and design an algorithm called rough entropy-based outlier detection (REOD) to find such outliers. Then, we regard intrusion activities as outliers and apply the REOD to intrusion detection, from which a novel approach for unsupervised intrusion detection is obtained. Experiments on several data sets demonstrate that the REOD performs well for outlier detection. In addition, compared with existing intrusion detection methods, the REOD can detect attacks with high detection rate and low false positive rate. Especially, the computational cost of the REOD is low, and it is suitable for intrusion detection on massive and high dimensional data.

Keywords: outlier detection；rough sets；roughness；rough entropy；unsupervised intrusion detection

0 引 言

粗糙集是处理不确定与不完备数据的有效工具,尤其是属性约简获得了大量研究[1-6]. 近年来,信息熵[7]被广泛应用于粗糙集,并出现了粗糙熵、知识熵和条件熵等新概念[3-4,8-9]. Qian等[1]提出组合熵,用来度量不完备信息系统的不确定性;Miao等[3]基于知识熵提出一种启发式约简方法[3];Wang等[4]利用条件信息熵进行决策表的约简;Liang等[9]提出了知识粗糙熵和粗糙集粗糙熵等新概念.

作为数据挖掘的一个重要研究方向,离群点检测关注于数据集中的一小部分对象,它们与其他对象存在显著的不同[10-12]. 目前,已出现了多种基于不同内容的离群点检测方法,如基于统计[10]、基于深度[11]、基于聚类[12]、基于密度[13]以及基于距离[14-15]的方法.

当前的离群点检测方法还存在一些问题,例如,不能有效处理不确定数据、计算开销大等. 针对这些问题,本文在粗糙集中定义一种新的粗糙熵,提出一种基于粗糙熵的离群点检测(rough entropy-based

收稿日期: 2018-10-06；**修回日期:** 2018-12-21.

基金项目: 国家自然科学基金项目(61402246, 61973180)；山东省自然科学基金项目(ZR2018MF007)；山东省重点研发计划项目(2018GGX101052).

责任编委: 王凌.

†通讯作者. E-mail: d_jw@163.com.

outlier detection, REOD)算法,并将REOD应用于无监督入侵检测.

近年来,无监督入侵检测引起大量学者的关注[16-24]. 无监督方法可以直接处理未标记的样本. Portnoy等[19]最早提出无监督入侵检测的思想,现有的无监督方法主要包括基于聚类的[19-21,24]、基于支持向量机的[17]、基于离群点检测[18]的方法. 离群点检测方法在入侵检测领域具有很大的应用潜力. 入侵行为与离群点非常相似,入侵行为相对于整个网络行为而言,是其中一小部分具有特殊属性的数据. 因此,只要将入侵行为看作是偏离于正常行为的一类离群点,就可以将离群点检测应用于无监督入侵检测[18].

现有的离群点检测方法不能有效处理不确定数据,而入侵检测所面对的是一个复杂的网络环境,具有高度的不确定性. 因此,将现有的离群点检测方法直接用于无监督入侵检测是不合适的,需要专门设计特定的离群点检测方法. 基于上述考虑,本文提出基于粗糙熵的离群点检测算法REOD,采取有效策略来处理不确定数据,从而可以在未标记数据上直接检测入侵. 实验表明,REOD的离群点检测性能要好于或等于已有方法,且其对入侵的检测效果也要好于已有的同类方法.

1 粗糙集的基本知识

在粗糙集中,信息表是一个四元组$\mathrm{IS}=(U,A,V,f)$. 其中:U和A分别是对象集和属性集;$V=\bigcup\limits_{a\in A}V_a$是所有属性论域的并;$f:U\times A\to V$是一个函数,使得对于任意$a\in A$和$x\in U$,有$f(x,a)\in V_a$.

给定信息表$\mathrm{IS}=(U,A,V,f)$,对任意$B\subseteq A$,定义由B所决定的不可区分关系$\mathrm{IND}(B)$为

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

$\mathrm{IND}(B)$是U上的一个等价关系,它将U划分成多个等价类,所有等价类的集合构成U的一个划分,记为$U/\mathrm{IND}(B)$. $\mathrm{IND}(B)$可看作是论域U上的一条知识[25]. 对任意$X\subseteq U$,X在知识$\mathrm{IND}(B)$下的粗糙度定义为

$$\rho_B(X)=1-|\underline{X}_B|/|\overline{X}_B|,$$

其中$\underline{X}_B$和$\overline{X}_B$分别表示X的B-下近似和B-上近似[25].

2 基于粗糙熵的离群点检测算法

2.1 离群点的定义

本文使用粗糙集中粗糙熵[8-9]的概念来检测离群点. 下面,首先给出一种新的粗糙集粗糙熵的定义.

定义1 给定信息表$\mathrm{IS}=(U,A,V,f)$,对任意$X\subseteq U$和$B\subseteq A$,令$X/\mathrm{IND}(B)=\{B_1,\dots,B_m\}$. 粗糙集$X$在$\mathrm{IND}(B)$下的粗糙熵$\mathrm{RE}_B(X)$被定义为

$$\mathrm{RE}_B(X)=\rho_B(X)+\mathrm{REK}_X(B).$$

其中:$\rho_B(X)$为X在$\mathrm{IND}(B)$下的粗糙度,$\mathrm{REK}_X(B)=-\sum\limits_{i=1}^{m}\dfrac{|B_i|}{|X|}\log_2\dfrac{1}{|B_i|}$为知识$\mathrm{IND}(B)$相对于$X$的粗糙熵(知识粗糙熵$\mathrm{REK}_X(B)$的具体定义见文献[9]).

粗糙集粗糙熵$\mathrm{RE}_B(X)$只是度量了粗糙集X的不确定性. 为了度量U中某个对象x对信息表IS不确定性的影响力,引入对象重要性的概念.

定义2(对象重要性) 给定信息表$\mathrm{IS}=(U,A,V,f)$,对任意$B\subseteq A$和$x\in U$,对象x在$\mathrm{IND}(B)$下的重要性$\mathrm{SIG}_B(x)$被定义为

$$\mathrm{SIG}_B(x)=\begin{cases}1-\dfrac{\mathrm{RE}_B(U)}{\mathrm{RE}_B(U-\{x\})},\ \mathrm{RE}_B(U-\{x\})>\mathrm{RE}_B(U);\\ 0,\ \text{otherwise}.\end{cases}$$

定义2的直观含义如下:当从U中去掉x后,如果$U-\{x\}$的粗糙熵比U的粗糙熵明显增加了,则表明$U-\{x\}$的粗糙度或知识$\mathrm{IND}(B)$的粗糙熵也明显增加了. 这样,就由x引发了IS不确定性的明显增加. 由于不确定性的明显增加是x带来的,x在$\mathrm{IND}(B)$下的重要性比较大.

接下来,通过考察对象的重要性信息来检测离群点. 离群点检测总是倾向于数据集中一小部分具有异常属性的对象[14],因此除了重要性信息之外,还将考虑对象的隶属信息. 为了获得对象的隶属信息,引入如下概念.

定义3(相对比重) 给定信息表$\mathrm{IS}=(U,A,V,f)$,对任意$B\subseteq A$,令$U/\mathrm{IND}(B)=\{B_1,\dots,B_m\}$. 对任意$x\in U$,等价类$[x]_B$的相对比重$\mathrm{RP}([x]_B)$定义为

$$\mathrm{RP}([x]_B)=\frac{|[x]_B|\times m}{|U|}\times\frac{|[x]_B|-\min+1}{\max-\min+2},$$

其中min和max分别表示集合$\{|B_1|,\dots,|B_m|\}$中取值最小和最大的元素.

$\mathrm{RP}([x]_B)$刻画的是x所在群体的相对规模,即相对于其他群体规模而言,x所在群体的规模.

定义4(属性序列) 给定信息表$\mathrm{IS}=(U,A,V,f)$,其中$A=\{a_1,\dots,a_p\}$. 对任意$1\leqslant j\leqslant p$,令$\mathrm{weight}(a_j)=\mathrm{REK}_U(A)-\mathrm{REK}_U(A-\{a_j\})$表示属性$a_j$的权重. 根据权重,对$A$中属性进行升序排列,从而得到属性序列$S=\langle a'_1,\dots,a'_p\rangle$,其中,对任意$1\leqslant j\leqslant p$,有$a'_j\in A$,并且对任意$1\leqslant j<p$,有$\mathrm{weight}(a'_j)\leqslant\mathrm{weight}(a'_{j+1})$.

定义5(属性子集序列) 给定信息表$\mathrm{IS}=(U,$

$A, V, f)$,其中$A = \{a_1, \ldots, a_p\}$. 令$S = \langle a'_1, \ldots, a'_p\rangle$为定义4中所给出的属性序列. $\mathrm{AS} = \langle A_1, \ldots, A_p\rangle$被称为IS中的一个属性子集序列,其中,对任意$1 \leqslant j \leqslant p$,有$A_j \subseteq A, A_1 = A, A_p = \{a'_p\}$,并且对任意$1 \leqslant j < p$,有$A_{j+1} = A_j - \{a'_j\}$.

为了度量对象的离群程度,下面给出“粗糙熵离群因子”这一概念[14-15].

定义6 (粗糙熵离群因子)　给定信息表$\mathrm{IS} = (U, A, V, f)$. 其中: $|U| = n, A = \{a_1, \ldots, a_p\}$. 令$\mathrm{AS} = \langle A_1, \ldots, A_p\rangle$为定义5中给出的属性子集序列,对任意$x \in U, x$的粗糙熵离群因子$\mathrm{REOF}(x)$定义为

$$\mathrm{REOF}(x) = \frac{\sum_{j=1}^{p}(1-(\mathrm{SIG}_{\{a_j\}}(x))^{\lambda}) \times (1-(\mathrm{RP}([x]_{\{a_j\}}))^{\lambda})}{2\times p} + \frac{\sum_{j=1}^{p}(1-(\mathrm{SIG}_{A_j}(x))^{\lambda}) \times (1-(\mathrm{RP}([x]_{A_j}))^{\lambda})}{2\times p},$$

其中$0 < \lambda \leqslant 1$是一个给定的参数.

定义7 (基于粗糙熵的离群点)　给定信息表$\mathrm{IS} = (U, A, V, f)$和阈值$\mu$,对于任意$x \in U$,如果$\mathrm{REOF}(x) > \mu$,则对象$x$被称为IS中的一个基于粗糙熵的离群点.

2.2　基于粗糙熵的离群点检测算法REOD

算法1

输入: 信息表$\mathrm{IS} = (U, A, V, f), |U| = n, A = \{a_1, \ldots, a_p\}$;参数$\lambda$、阈值$\mu$.

输出: U中的离群点以及每个对象的离群因子.

step 1: 采用计数排序的方法,计算划分$U/\mathrm{IND}(A)$[5].

step 2: 根据$U/\mathrm{IND}(A)$,计算知识粗糙熵$\mathrm{REK}_U(A)$.

step 3: 对任意$a_i \in A, 1 \leqslant i \leqslant p$,循环执行:

step 3.1: 采用计数排序方法,计算$U/\mathrm{IND}(A - \{a_i\})$;

step 3.2: 计算知识粗糙熵$\mathrm{REK}_U(A - \{a_i\})$;

step 3.3: 计算属性a_i的权重$\mathrm{weight}(a_i)$.

step 4: 构建属性序列S和属性子集序列$\mathrm{AS} = \langle A_1, \ldots, A_p\rangle$.

step 5: 对任意$1 \leqslant i \leqslant p$, 循环执行:

step 5.1: 计算划分$U/\mathrm{IND}(\{a_i\})$和$U/\mathrm{IND}(A_i)$;

step 5.2: 计算知识粗糙熵$\mathrm{REK}_U(\{a_i\})$和$\mathrm{REK}_U(A_i)$;

step 5.3: 计算粗糙集粗糙熵$\mathrm{RE}_{\{a_i\}}(U)$和$\mathrm{RE}_{A_i}(U)$.

step 6: 对任意$x \in U$,循环执行:

step 6.1: 对任意$1 \leqslant i \leqslant p$,循环执行:

step 6.1.1: 计算$\mathrm{REK}_{U-\{x\}}(\{a_i\})$、$\mathrm{REK}_{U-\{x\}}(A_i)$;

step 6.1.2: 计算$\mathrm{RE}_{\{a_i\}}(U-\{x\})$、$\mathrm{RE}_{A_i}(U-\{x\})$;

step 6.1.3: 计算x的重要性$\mathrm{SIG}_{\{a_i\}}(x)$、$\mathrm{SIG}_{A_i}(x)$;

step 6.1.4: 计算相对比重$\mathrm{RP}([x]_{\{a_i\}})$、$\mathrm{RP}([x]_{A_i})$.

step 6.2: 计算对象x的粗糙熵离群因子$\mathrm{REOF}(x)$.

step 7: 根据离群因子,对所有对象进行降序排列.

step 8: 选择离群因子大于μ的对象作为离群点.

step 9: 算法结束,返回所有离群点及对象的离群因子.

对任意$B \subseteq A$,算法1采用了一种预先对U中对象进行计数排序,然后再求划分$U/\mathrm{IND}(B)$的方法[5]. 在最坏的情况下,算法1的时间复杂度为$O(|A|^2 \times |U|)$,空间复杂度为$O(|A| \times (|A| + |U|))$.

2.3　实验结果

下面,通过实验来验证REOD的性能. 首先,在UCI数据集Lymphography[26]上比较REOD、KNN[27]和基于距离的离群点检测方法(简称DIS)[14]的性能;其次,在数据集Breast Cancer上[26],比较REOD、KNN、RNN[28]和DIS这4种方法的性能,其中RNN的实验结果可参见文献[28].

对于KNN,设置参数k为5[27]. 对于DIS,具体实验细节请参见文献[14]. 另外,对于REOD,通过多次实验尝试来确定参数λ的取值. 具体而言,首先为λ设置一个初始经验值,并验证相应的离群点检测结果;然后,不断调整λ的值并验证调整后的实验结果,直到得到满意的结果;最终,λ被设置为0.55.

实验采用文献[29]中提出的评价指标体系来评测各个方法的性能. 首先,在Lymphography上进行实验. 该数据集包含148个对象和19个属性,共6个离群点. 具体实验结果如表1所示.

表1　参数设置

离群程度值前k%的对象(对象个数)	属于离群点的对象个数(覆盖率/%)		
	REOD	DIS	KNN
3% (4)	4(67)	4(67)	3(50)
4% (6)	5(83)	5(83)	4(67)
5% (7)	6(100)	5(83)	4(67)
6% (9)	6(100)	6(100)	4(67)
8% (12)	6(100)	6(100)	5(83)
10% (15)	6(100)	6(100)	6(100)

表1中: “属于离群点的对象个数”是指在由某个方法所计算出的离群程度值排在前k%的对象中,真正的离群点个数;“覆盖率”是指目前已检测出的离群点占整个离群点的比例[14-15]. 从表1可以看出,REOD的性能明显要好于DIS和KNN.

其次,在Breast Cancer上进行实验. 该数据集包含699个对象和9个属性. 为了获得一个不平衡的数

据集,删除了其中一些“malignant”对象(即属于恶性肿瘤的样本). 最终的数据集包含39个“malignant”和444个“benign”对象(属于良性肿瘤的样本). 另外,对该数据集中的连续型属性进行了离散化处理[28]. Breast Cancer上的实验结果如表2所示.

表 2 Breast Cancer上的实验结果

离群程度值前k%的对象(对象个数)	属于离群点的对象个数(覆盖率/%)			
	REOD	DIS	RNN	KNN
1%(4)	4(10)	4(10)	3(8)	4(10)
2%(8)	8(21)	5(13)	6(15)	8(21)
4%(16)	16(41)	11(28)	11(28)	16(41)
6%(24)	21(54)	18(46)	18(46)	20(51)
8%(32)	28(72)	24(62)	25(64)	27(69)
10%(40)	33(85)	29(74)	30(77)	32(82)
12%(48)	37(95)	36(92)	35(90)	37(95)
14%(56)	39(100)	39(100)	36(92)	39(100)
18%(72)	39(100)	39(100)	38(97)	39(100)
28%(112)	39(100)	39(100)	39(100)	39(100)

3 REOD算法在无监督入侵检测中的应用

3.1 设计方案

本节将REOD算法应用于无监督入侵检测,进而得到一种新的无监督入侵检测方法. 该方法无需对入侵行为或正常行为建模,而是通过检测离群点的方式直接检测入侵,其基本流程如图1所示.

图 1 方法的基本流程

如图1所示,本文提出的无监督入侵检测方法的基本流程可分为4个步骤: 1)采集网络数据包或日志文件,并进行解码、过滤、统计、分析; 2)对数据进行预处理,包括补齐、离散化、特征选择等; 3)采用REOD对预处理后的数据进行离群点检测; 4)将上一步检测出的离群点标记为入侵,并响应入侵.

3.2 实 验

下面,通过KDD Cup 99数据集来验证REOD的入侵检测性能. 该数据集包含约490万条记录,所有攻击类型被分成DOS、R2L、U2R和Probe四大类[26]. 该数据集过于庞大,因此只选取了其中有代表性的一个子集“10%-KDD”(包含494 021条记录)[26].

3.2.1 数据预处理

10%-KDD有41个属性,其中包含了一些连续型属性[26]. 粗糙集更适合于处理离散型属性,因此需要对连续型属性进行离散化[25],采用文献[30]中提出的离散化算法SMDNS. 在离散化时,由于部分连续型属性的取值个数非常少,没有必要进行离散化,因此,只针对其中的26个属性(即序号为1、5、6、10、13、16、17、23～41的属性)进行离散化.

另外,不是所有属性对最终的入侵检测结果都有相同的影响,对此,本文进行了特征选择. 最终,选择17个属性(即序号为3、4、5、8、10、14、23～31、40、41的属性)用于入侵检测.

3.2.2 实验过程与结果

实验过程分为如下两个阶段.

1) 针对DOS、R2L、U2R和Probe,分别测试REOD对每类攻击的检测效果. 从10%-KDD中提取出全部97 278条正常记录,构成数据集N. 另外,对每类攻击,依据不同比例随机不放回地抽取出部分攻击记录,构成数据集D、R、U、P,其中D、R、U、P分别存放DOS、R2L、U2R和Probe类型的攻击. 每类攻击的抽取比例与记录数见表3.

表 3 每类攻击的抽取比例及数目

攻击类型	抽取比例/%	抽取记录数
DOS	0.311	1 217
R2L	50	563
U2R	100	52
Probe	24.84	1 020

对于每类攻击,分别将N与相应的攻击数据集合并,形成4个测试集: $N\bigcup D$、$N\bigcup R$、$N\bigcup U$和$N\bigcup P$. 对于每个测试集,先进行离散化与特征选择,然后运行REOD,以验证REOD对该类攻击的检测能力. 为了避免随机性给实验结果带来的不稳定性,上述实验重复100次,最终结果取100次的平均.

2) 测试REOD对所有攻击的检测效果. 从10%-KDD中提取出全部97 278条正常记录,构成数据集N. 另外,从396 743条攻击记录中(包含各种类型的攻击)依据一定的比例(0.504%)随机不放回地抽取2 000条攻击记录,构成数据集I. 将N与I合并,得到测试集$N\bigcup I$. 对于$N\bigcup I$,先进行离散化与特征选择,然后运行REOD,以验证REOD对混合攻击的检测能力. 同样,上述实验重复100次,最终结果取100次的平均.

为了便于比较,下面将狄利克雷混合模型[16]、遗传聚类[21]、度量学习[22]、遗传算法[23]、离群点检测[18]以及无监督聚类[24]共计6种入侵检测方法的实验结果与REOD的结果列在一起,如表4所示.

表4中,DR(detection rate)表示检测率,FPR(false positive rate)表示误报率. 从表4可以看出,当使用REOD检测入侵时,对于大部分入侵网络连接,其离群程度值明显高于正常的网络连接. 因此,REOD可以有效区分正常行为与入侵行为.

通过对比各种方法的结果,可以看出: REOD对于DOS和Probe攻击的检测效果非常好,而对于R2L和U2R,REOD的检测效果也明显好于其他方法. 无

表 4　入侵检测结果对比

入侵检测方法	DOS攻击		R2L攻击		U2R攻击		Probe攻击		所有攻击	
	DR/%	FPR/%	DR/%	FPR/%	DR/%	FPR/%	DR/%	FPR/%	DR/%	FPR/%
狄利克雷混合模型	87.6	2.1	62.1	3.5	71.5	1.1	89.8	1.9	86.4	1.8
遗传聚类	56	/	66	/	78	/	44	/	59.6	/
度量学习	91.2	1.7	64.3	3.8	81.9	1.2	96.3	1.8	92.1	1.6
遗传算法	94.0	/	68.1	/	76.3	/	90.4	/	93.2	1
无监督聚类	59.5	/	40.2	/	69.1	/	78.6	/	61.9	/
离群点检测	94.6	/	65.5	/	73.2	/	95.5	/	95.0	1
REOD算法	99.2	0.9	75.4	3.2	88.1	0.7	95.9	1.5	97.2	0.9

论是检测哪一种类型的攻击,REOD的检测效果明显要好于以下5种方法:狄利克雷混合模型、遗传聚类、遗传算法、无监督聚类以及离群点检测方法. 与度量学习方法相比,REOD能够更好地检测出DOS、R2L和U2R这3类攻击,只是对Probe的检测率稍低一些. 另外,REOD在误报率上表现更好. 因此,总体而言,REOD的表现仍然是最优的.

3.2.3　计算性能分析

为了评估REOD的计算性能,在一台内存为8 GB、CPU为3.4 GHz的PC机上统计不同方法在前面生成的两个测试集上的运行时间. 其中:测试集$N \bigcup D$包含98 495个对象和41个属性,测试集$N \bigcup I$包含99 278个对象和41个属性. 所有方法在这两个测试集上的运行时间(单位: s)如表5所示.

表 5　不同方法在两个测试集上的运行时间

入侵检测方法	$N \bigcup D$	$N \bigcup I$
狄利克雷混合模型	206	209
遗传聚类	6 351	6 419
度量学习	2 865	2 903
遗传算法	25 194	25 549
无监督聚类	68	70
离群点检测	2 412	2 447
REOD	7	7

进一步,从完整的KDD Cup 99数据集中抽取全部972 781条正常记录,并随机抽取2万条攻击记录,从而得到第3个测试集(包含约一百万条记录). 该测试集上各方法的运行时间见图2.

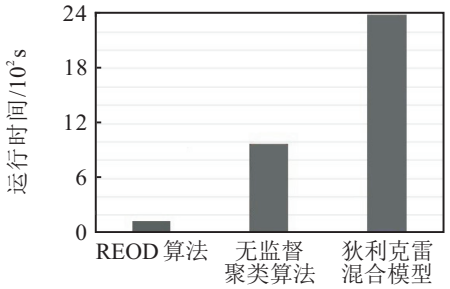

图 2　不同方法在第3个测试集上的运行时间

从表5以及图2可以看出,REOD的计算性能明显优于其他方法.

4　结　论

本文利用粗糙集中的粗糙熵来进行离群点检测,并提出离群点检测算法REOD. 通过将入侵行为看作是离群点,进一步将REOD应用于无监督入侵检测. 该方法无需对入侵行为或正常行为建模,而是采用离群点检测的策略直接在未标记的数据上检测入侵,避免了有监督方法的问题. 实验表明,REOD可以有效地检测出入侵行为,特别是它的计算开销比较小,这对于在实际环境中运行的入侵检测系统非常重要.

下一步工作将考虑结合其他机器学习算法来对本文方法进行改进,进一步提高其对R2L和U2R攻击的检测率. 另外,计划利用扩展的粗糙集模型来改进本文方法,例如利用邻域粗糙集[2,31]来检测离群点,从而不需要离散化就可以直接处理连续型属性.

参考文献(References)

[1] Qian Y H, Liang J Y, Wang F. A new method for measuring the uncertainty in incomplete information systems[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2009, 17(6): 855-880.

[2] Dai J H, Hu Q H, Hu H, et al. Neighbor inconsistent pair selection for attribute reduction by rough set approach[J]. IEEE Transactions on Fuzzy Systems, 2018, 26(2): 937-950.

[3] Miao D Q, Hu G R. An heuristic algorithm of knowledge reduction[J]. Computer Research and Development, 1999, 36(6): 681-684.

[4] Wang G Y, Yu H, Yang D C. Decision table reduction based on conditional information entropy[J]. Chinese Journal of Computers, 2002, 25(7): 759-766.

[5] Xu Z Y, Liu Z P, Yang B R, et al. A quick attribute reduction algorithm with complexity of $\max(O(|C||U|), O(|C|^2|U/C|))$[J]. Chinese Journal of Computers, 2006, 29(3): 391-399.

[6] 陈迎春, 李鸥, 孙昱. 基于聚类离散化和变精度邻域熵的属性约简[J]. 控制与决策, 2018, 33(8): 1407-1414. (Chen Y C, Li O, Sun Y. Attribute reduction based on clustering discretization and variable precision neighborhood entropy[J]. Control and Decision, 2018, 33(8): 1407-1414.)

[7] Shannon C E. The mathematical theory of

communication[J]. Bell System Technical Journal, 1948, 27(3/4): 373-423.

[8] Beaubouef T, Petry F E, Arora G. Information-theoretic measures of uncertainty for rough sets and rough relational databases[J]. Information Sciences, 1998, 109: 535-563.

[9] Liang J Y, Shi Z Z, Li D Y, et al. Information entropy, rough entropy and knowledge granularity in incomplete information systems[J]. International Journal of General Systems, 2006, 35(6): 641-654.

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

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

[12] 耿志强, 姬威, 韩永明, 等. 基于维度最大熵数据流聚类的异常检测方法[J]. 控制与决策, 2016, 31(2): 343-348.
(Geng Z Q, Ji W, Han Y M, et al. Data stream clustering algorithm based on the maximum entropy of data dimension and its applications for anomaly detection[J]. Control and Decision, 2016, 31(2): 343-348.)

[13] 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. Dallas: ACM Press, 2000: 93-104.

[14] Jiang F, Sui Y F, Cao C G. A hybrid approach to outlier detection based on boundary region[J]. Pattern Recognition Letter, 2011, 32(14): 1860-1870.

[15] 江峰, 杜军威, 眭跃飞, 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, Sui Y F, et al. Outlier detection based on boundary and distance[J]. Chinese Journal of Electronics, 2010, 38(3): 700-705.)

[16] Singh J P, Bouguila N. Intrusion detection using unsupervised approach[C]. Proceedings of the International EAI Conference on Emerging Technologies for Developing Countries. Morocco: Springer, 2017: 192-201.

[17] Nguyen B V. An application of support vector machines to anomaly detection[R]. Athens: Research in Computer Science-Support Vector Machine, 2002.

[18] Zhang J, Zulkernine M. Anomaly based network intrusion detection with unsupervised outlier detection[C]. IEEE International Conference on Communications. Istanbul: IEEE, 2006: 2388-2393.

[19] Portnoy L, Eskin E, Stolfo S J. Intrusion detection with unlabeled data using clustering[C]. Proceedings of the ACM Workshop on Data Mining Applied to Security. Philadelphia: ACM Press, 2001: 113-125.

[20] Eskin E, Arnold A, Prerau M, et al. A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data[C]. Applications of Data Mining in Computer Security. Boston: Springer, 2002: 78-99.

[21] Liu Y G, Chen K F, Liao X F, et al. A genetic clustering method for intrusion detection[J]. Pattern Recognition, 2004, 37(5): 927-942.

[22] Aliakbarisani R, Ghasemi A, Felix Wu S. A data-driven metric learning-based scheme for unsupervised network anomaly detection[J]. Computers& Electrical Engineering, 2019, 73: 71-83.

[23] 张凤斌, 杨永田, 江子扬. 遗传算法在基于网络异常的入侵检测中的应用[J]. 电子学报, 2004, 32(5): 875-877.
(Zhang F B, Yang Y T, Jiang Z Y. Genetic algorithms in intrusion detection based on network anomaly[J]. Chinese Journal of Electronics, 2004, 32(5): 875-877.)

[24] 罗敏, 王丽娜, 张焕国. 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, Zhang H G. An unsupervised clustering-based intrusion detection method[J]. Chinese Journal of Electronics, 2003, 31(11): 1713-1716.)

[25] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Sciences, 1982, 11: 341-356.

[26] Bache K, Lichman M. The UCI machine learning repository[EB/OL]. (2013-12-23)[2018-10-06]. http://archive.ics.uci.edu/ml.

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

[28] Harkins S, 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. Aix-en-Provence: Springer-Verlag, 2002: 170-180.

[29] Aggarwal C C, Yu P S. Outlier detection for high dimensional data[C]. Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data. Santa Barbara: ACM Press, 2001: 37-46.

[30] Jiang F, Sui Y F. A novel approach for discretization of continuous attributes in rough set theory[J]. Knowledge-Based Systems, 2015, 73: 324-334.

[31] 黄恒秋, 曾玲, 黎利辉. 混合值不完备系统的双邻域粗糙集分类方法[J]. 控制与决策, 2018, 33(7): 1207-1214.
(Huang H Q, Zeng L, Li L H. Double-neighborhood rough set classification method in incomplete decision system with hybrid value[J]. Control and Decision, 2018, 33(7): 1207-1214.)

作者简介

江峰(1978−), 男, 副教授, 博士, 从事机器学习、粗糙集等研究, E-mail: jiangfeng@qust.edu.cn;

王凯郦(1994−), 女, 硕士生, 从事数据挖掘、机器学习的研究, E-mail: 707444834@qq.com;

于旭(1982−), 男, 副教授, 博士, 从事机器学习、推荐系统等研究, E-mail: yuxu0532@163.com;

眭跃飞(1963−), 男, 研究员, 博士, 从事大规模知识处理的理论基础、知识表示的研究, E-mail: 14269672@qq.com;

杜军威(1974−), 男, 教授, 博士, 从事机器学习、知识图谱与知识工程等研究, E-mail: d_jw@163.com.

(责任编辑: 齐 霁)