网络入侵检测的GEP规则提取算法研究

唐 菀[1,2] 曹 阳[1] 杨喜敏[2,3] 覃 俊[2]

(武汉大学电子信息学院软件工程国家重点实验室 武汉 430070)[1]

(中南民族大学计算机科学学院 武汉 430074)[2] (华中科技大学计算机科学与技术学院 武汉 430074)[3]

摘 要 针对基于机器学习网络入侵检测存在的未知攻击检测率低、规则多而复杂导致检测效率不高等问题,提出了基于约束的基因表达式编程(GEP)规则提取算法(CGREA)。用GEP模式表示入侵检测规则,定义了约束文法对规则个体进行约束,以满足规则的充分性和封闭性。CGREA算法限定GEP规则基因头部各类符号的随机选择数目比例,并采用精英策略以保证算法收敛性。用KDD CUP'99数据集对CGREA算法提取的入侵检测规则进行评估,总攻击检测率为91.36%,其中有3种未知攻击的检测率超过88%。结果表明,CGREA算法能在较小种群和有限代数内提取出简单而有效的规则,未知攻击检测率和检测性能也得到提高。

关键词 网络入侵检测,基因表达式编程,规则提取,约束文法,精英策略

中图法分类号 TP393 **文献标识码** A

Study on GEP Rule Extraction Algorithm for Network Intrusion Detection

TANG Wan[1,2] CAO Yang[1] YANG Xi-min[2,3] QIN Jun[2]

(State Key Laboratory of Software Engineering, School of Electronic Information, Wuhan University, Wuhan 430070, China)[1]

(College of Computer Science, South-Central University for Nationalities, Wuhan 430074, China)[2]

(College of Computer Science & Technology, Huazhong University of Science & Technology, Wuhan 430074, China)[3]

Abstract Network intrusion detection based on machine learning suffers from the problems of low detection ratio for unknown intrusion and low detection efficiency due to many complex rules. To solve these problems, a constraint-based gene expression programming (GEP) rule extraction algorithm (CGREA) was proposed. The intrusion detection rules were represented based on GEP model, and a constraint grammar was defined to guarantee the rules closeness and adequacy. It restricted the ratio of randomly selecting various symbols in the gene head of GEP rules, and used the elitist strategy to guarantee convergence. The KDD CUP'99 DATA Set was used for evaluation the intrusion detection rules auto-extracted by CGREA. A 91% probability of detection was achieved, and three unknown attacks' probabilities of detection were more than 88%. These results indicate that the intrusion detection rules that extracted by CGREA are effective, simple, and capable of detecting unknown intrusions. Moreover, the efficiency of rule generation and detection is improved.

Keywords Network intrusion detection, GEP (gene expression programming), Rule extraction, Constraint grammar, Elitist strategy

1 引言

入侵检测(Intrusion Detection, ID)是当前计算机网络安全领域的一个研究热点,主要有误用检测和异常检测两种检测方法,机器学习是异常检测的主要方法之一。机器学习模型依据提供给定特征的网络数据,能部分或全部自动地提取其中隐藏的、能识别正常行为特征的规则,并使用它们来判断一个新网络事件是否正常。目前用于入侵检测规则获取的机器学习方法主要有支持向量机(SVM)[1]、决策树[2-4]、进化算法[5,6]等。决策树和SVM检测性能较好,但SVM适用于小样本学习,对大量的网络流量数据还需要事先生成多层次的训练小样本[1],而决策树方法生成的规则决策树数目多且需要规则剪枝等操作[2]。文献[5]结合聚类分析和进化算法建立入侵检测模型,利用进化算法对聚类准则函数进行优化,但训练集数量大小会直接影响检测结果。Abraham等将基因表达式编程(Gene Expression Programming, GEP)、线形遗传编程(LGP)、多表达式编程(MEP)均看作是改进的遗传编程(GP)[6],也用来自动提取入侵检测规则,但未讨论对新攻击

到稿日期:2008-12-09 返修日期:2009-03-09 本文受国家重点基础研究发展计划(2004CB318203),国家自然科学基金(60603008),湖北省自然科学基金(BZY07008)资助。

唐 菀(1974-),女,博士研究生,副教授,CCF会员,主要研究方向为高速网络安全等,E-mail:tangwan_y@hotmail.com;曹 阳(1943-),男,教授,博士生导师,主要研究方向为高速网络关键技术等;杨喜敏(1972-),男,博士研究生,主要研究方向为存储系统性能分析等;覃 俊(1968-),女,博士,教授,主要研究方向为智能算法与信息挖掘等。

的检测能力，且由于测试数据量小，无法与其他方法进行有效比较。因此，机器学习应用到实际入侵检测系统中仍然存在未知攻击检测能力低、运算量大、检测效率不高等诸多问题。

GEP 于 2001 年由 Ferreira 提出[6]，集成了遗传算法(GA)简单快捷的定长线形编码和 GP 灵活多样的属性结构等特点，用简单编码解决复杂问题。与 GP 相比，GEP 能避免维度爆炸问题，遗传操作更加灵活多样。国内外不少研究人员已将 GA 和 GP 用于入侵检测规则表示及提取，但除文献[6]外，鲜有将 GEP 应用到网络入侵检测研究的文献。在解决具体实际问题时，为了能充分描述问题解，需要输入输出变量是不同类型的多种函数，但由于 GEP 个体的封闭性无法满足，导致无效个体的生成。本文将利用 GEP 的优势，采用 GEP 约束文法获得符合封闭性的入侵检测规则，并提出基于约束的 GEP 规则提取算法 CGREA，来实现自动提取高效入侵检测规则，同时提高对未知攻击的检测能力。

2 基于约束的 GEP 规则提取算法(CGREA)

2.1 GEP 规则约束文法

GEP 个体由固定长度符号串表示，进化过程采用交叉、变异等遗传算子不断更新并传送到下一代种群。GEP 个体表达方式又称为 K 表达式，一个 K 表达式可对应一个表达式树。K 表达式是 GEP 基因型表示，便于遗传操作。对个体解码得到的表达式树是 GEP 显性表示，便于解析语义。评价适应度值时，将个体从串形式转换成表达式树，使搜索空间与问题解空间分离，并保证输出的有效性[6]。

作为表达式树的基本构成元素，表示规则的 GEP 个体的函数集和终结符集(包括属性集和常量集)要满足封闭性和充分性。封闭性是指每个函数都能以任何终结符或任何函数的结果作为自变量，保证产生的个体语法的正确性。充分性指函数和终结符的组合可以生成问题解[8]。为了最大限度实现充分性，需要输入输出变量是不同类型的多种函数，但这又无法满足封闭性，不能保证 GEP 模式获得的个体在解决具体实际问题时完全有效。例如，当函数集为{+，−，and，or，>，<，=}、终结符集为{5，a，b，c，d，e}时，某次变异操作后获得一个长度为 12 的 GEP 单基因个体为

$$\text{or}.+.>.a.b.>.5.c.d.a.a.d \tag{1}$$

式(1)对应的表达式树如图 1 所示。

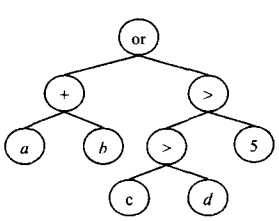

图 1 式(1)对应的 GEP 表达式树

将图 1 中的表达式树采用宽度优先方式遍历，得到

$$(a+b)\ \text{or}\ ((c>d)>5) \tag{2}$$

虽然式(2)符合 GEP 模式定义，但不满足封闭性且无效。本文采用前期研究提出的约束文法 G 来保证 GEP 规则的封闭性[9]。下面是 G 的形式化描述

文法 $G=(E,T,E_L,P)$

其中，基本符号串类型集合 $E=\{E_L,E_{LR},E_R,E_A,E_T\}$，终结符集 $T=\{l,r,a,v,c\}$，$l\in L$，$r\in R$，$a\in A$，L 为逻辑运算符集，R 为关系运算符集，A 为算术运算符集，L，R，A 均是函数集子集；$v\in A_s$，$c\in C_s$，A_s 和 C_s 分别为属性集和常量集，且 v 的值是经过数值化预处理的。

产生式集合 P 定义为

$E_L\rightarrow lE_{LR}^k\mid rE_{LR}^k$；$k$ 为 l 或 r 的目数

$E_{LR}\rightarrow E_R\mid E_L$

$E_R\rightarrow rE_A^2\mid rE_AE_T\mid rvE_A\mid rvE_T$

$E_A\rightarrow avE_T^{n-1}$；$n$ 为 a 的目数

$E_T\rightarrow v\mid c$

定义 1 集合 $L_{ang}(G)=\{w\in\{l,r,a,v,c\}^*:E_L\Rightarrow w\}$ 是满足文法 G 的 GEP 规则个体集。

定理 1 集合 $L_{ang}(G)$ 中的个体均满足封闭性。

证明：E_L 是规则个体的具体表示，而且对应的表达式输出是布尔值。当值为 1，表示匹配识别成功，否则失败。当 E_L 由逻辑运算符开始时，要求 E_{LR} 的输出是布尔值；当 E_L 由比较运算符开始时，E_{LR} 可以是属性值、常量值和布尔值。根据 P 的定义，每一个 E_L 串最初是从单个符号的 E_T 开始，然后依次生成 E_A，E_R，E_{LR}。从 E_R 的定义可知，它的输出是一个布尔值，而 E_L 的输出也是布尔值，因此 E_{LR} 的输出一定是个布尔值，满足了 E_L 的输入数据类型要求(证毕)。

2.2 算法描述

基于规则约束文法 G，本节提出一个基于约束的 GEP 规则提取算法(Constraint-based GEP-Rule Extraction Algorithm，CGREA)，采用文献[10]给出的适应度函数来确定每个个体的适应度值。

Step 1 对训练数据进行简单归一化处理。

Step 2 生成 GEP 规则 r_i 作为规则种群 S_{gen} 的初始个体，且 $r_i\in L_{ang}(G)$，$i=1,\cdots,p_n$，p_n 是 S_{gen} 的规模。r_i 基因头的函数符、属性符和常量符按照 $p_{fun}:p_{var}:(1-p_{fun}-p_{var})$ 比例均匀选择。

Step 3 建立临时规则种群 S_{tmp} 和最优规则个体种群 S_{best}，S_{tmp} 和 S_{best} 种群大小分别为 p_n 和 p_{best}，且 $p_{best}\leqslant p_n$。

Step 4 计算 S_{gen} 中所有个体的适应度值。更新 S_{best}，保证其中的个体是($S_{gen}\cup S_{best}$)中适应度值最高的前 p_{best} 个体。如果 S_{best} 有连续 g_m 代未更新，转 Step 8 继续。

Step 5 从 S_{gen} 中选择两个父个体 r_{p1}，r_{p2}。按照 p_{corss}，p_{muta} 的概率对 r_{p1} 和 r_{p2} 进行单点交叉和均匀变异，生成两个子个体 r_{o1} 和 r_{o2}。

Step 6 如果 $r_{oj}\notin L_{ang}(G)$，$j\in\{1,2\}$，重复对 r_{oj} 按照概率 p_{muta} 执行均匀变异，直到 $r_{oj}\in L_{ang}(G)$。如果重复次数等于 $c_{\max}$ 时仍然 $r_{oj}\notin L_{ang}(G)$，转 Step 5 继续。

Step 7 将 r_{o1} 和 r_{o2} 添加到 S_{tmp}。如果 S_{tmp} 个体数目$<p_n$，执行 Step 4，否则 $S_{gen}=S_{tmp}$。如果当前进化代数 g 没有超过最大代数 $g_{\max}$，则 g 增加 1 后转 Step4 继续。

Step 8 对 S_{best} 中个体执行 Step1 逆操作，得出有效入侵检测规则集。

2.3 算法分析

与基本 GEP 进化过程比较，CGREA 算法的以下改进使其更加适合于入侵检测规则的提取。

(1) 对训练数据进行简单归一化，使终结符集中元素均能采用同样变异方式，比离散化和模糊化更易实现。

(2) 随机生成新规则个体时，限定基因头部的函数符、属

性符、常量符的随机选择数目比例，避免在等概率下由于GEP属性集元素数目过多、函数集和常量集元素少，造成随机生成的初始规则大多为无效规则。

(3) 在进化过程中判断规则约束，使函数集可包含多种类型函数，提高规则通用性并适用于入侵检测。

(4) 采用精英策略，用最优规则集保存进化过程中适应度值最好的规则个体子集，防止全局优良个体在变异和交叉遗传操作中丢失。

CGREA算法复杂度为 $O(n^2)$，与基本GEP进化过程相同。预处理、新规则符号比例限制和最优规则集使用增加了一定的计算时间和存储空间。对规则的约束虽然增加了判断及反复生成的处理，但缩小了个体范围。

2.4 算法收敛性

CGREA算法进化过程与GA的基本相同，均属于多目标优化进化算法(multi-objective evolutionary algorithms, MOEA)。下面按照MOEA收敛性分析方法分析它的收敛性[8,11]。

令 G_t 是 S_{best} 第 t 代的种群，$F_t(G_t)$ 是其相应的适应度值(即目标空间)集合，F^* 是有限目标空间集合的极小元集，并假设 $|S_{best}|>|F^*|$。

首先，由于算法中的参数不随时间 t 改变，因此有

$$P\{G_{t+1}=j|G_t=i,G_{t-1}=i_{t-1},\cdots,G_0=i_0\}=P\{G_{t+1}=j|G_t=i\}=P_{i,j}$$

根据有限马尔可夫链定义，种群序列 (G_t) 的转移矩阵 P 与时间 t 无关。本算法中的变换算子(均匀变异和单点交叉)交替在每代中执行，每个个体通过当前代的进化，存在变化成为任意一个个体的可能性，因此它的转移矩阵是正的。采用确定性的选择算子，每次淘汰 S_{best} 中最差个体，选择父个体采用轮盘赌选择策略，因此选择算子的转移矩阵是可列、可容的。由于进化算法中的转移矩阵是变换算子和选择算子的转移矩阵的积，可知种群序列 (G_t) 是具有正转移矩阵的马尔可夫链。

其次，算法将当前代种群及上代 S_{best} 中适应度值最高的前 p_{best} 个规则个体组成新一代 S_{best} 个体，保证了在历代中获得的最优个体一旦进入 S_{best} 中就不再被淘汰，而 S_{best} 中的非最优个体在有限步内必将被取代，以概率1被淘汰。

综上可证CGREA算法以概率1收敛于极小元集 F^*。

3 实验

3.1 实验数据集

KDD CUP'99数据集是研究基于机器学习入侵检测的权威实验数据集[12]，它将网络数据分为Normal和4类入侵攻击(DoS, Probing, R2L, U2R)。本文选择其中的训练集(kddcup.data_10_percent)和测试集(corrected)作为实验数据集，并随机从训练集中选择数据作为训练子集。KDD数据集及训练子集分布如表1和表2所列。

表1 KDD CUP'99数据集中正常与攻击记录的分布

数据类型	训练集数据数及所占比例	测试集数据数及所占比例	测试集未知攻击数及在类别中所占比例
Normal	97278,19.69%	60593,19.48%	
Probing	4107,0.83%	4166,1.35%	1789,42.94% (mscan,saint)
DoS	391458,79.24%	229853,73.90%	6555,2.85% (apache2, mailbomb,processtable, udpstorm)
U2R	52,0.01%	228,0.07%	189,82.89% (xterm,ps, sqlattack)
R2L	1126,0.23%	16189,5.20%	10196,62.98% (httptunnel, xlock,xsnoop,named,sendmail, snmpgetattack,snmpguess, worm)

表2 训练子集中正常与攻击记录的分布

训练子集	normal	DoS	Probing	R2L	U2R
training_attack	4000,20.00%	13715,68.58%	1107,5.53%	1126,5.63%	52,0.26%
training_normal	10000,33.33%	14715,49.05%	4107,13.69%	1126,3.75%	52,0.17%

3.2 数据预处理和GEP参数

对训练数据进行简单预处理。首先将protocol_type, service, flag属性值转换为整数类型，然后对取值范围大于1的所有属性值分别按照下式归一化

$$x=(x-x_{\min})/(x_{\max}-x_{\min}),x\in[x_{\min},x_{\max}] \quad (3)$$

其中，x 为处理前的值，x' 为处理后的归一化值，$x_{\min}$ 和 $x_{\max}$ 分别是最小和最大属性值。

主要GEP参数及取值如表3所列，其中属性集元素a1—a41分别对应KDD CUP'99数据的41个属性。

表3 GEP主要参数表

参数	取值	参数	取值
函数集s	{and,or,not,>,>=,<,<=,==,!=}	基因头长	8
属性集s	{a1,a2,…,a41}	种群规模	100
常量集	[0,1]之间的随机常数	最优种群规模	20
选择算法	轮盘赌	P_{corss}	0.2
P_{avrt}	0.28	P_{muta}	0.95
P_{fun}	0.60	P_{best}	0.20

3.3 实验与结果分析

3.3.1 实验1——规则有效性

使用表2中的training_attack和training_normal训练子集分别生成攻击规则和Normal规则。将生成的每类规则中适应度值最高的1条规则作为检测规则，表4中列出了这5条规则及其进化生成代数。

表4 基于CGREA算法的各类最优入侵检测规则

规则	描述	生成代数
rule0	IF ((diff_srv_rate<=0.0382091738283634) and (dst_host_same_src_port_rate<0.924649775028229)) THEN is Normal	10
rule1	IF ((dst_host_diff_srv_rate>0.137455374002457) and (protocol_type=icmp or protocol_type=tcp)) THEN is Probing	3
rule2	IF ((protocol_type=icmp or protocol_type=tcp) and (count>=50)) THEN is DoS	19
rule3	IF ((duration<>2343.630205039) and (root_shell>0)) THEN is U2R	0
rule4	IF ((dst_bytes>1201907) or (is_guest_login>0)) THEN is R2L	13

文献[6]也将GEP用于入侵检测规则自动提取。生成每一种类型规则时使用的GEP参数有所不同，最大代数为500或800，基因个数为12或14。本文采用单基因个体且获得的最优规则的最大生成代数是19，比文献[6]的进化代数少。从表4中可以看到，本文提出的CGREA算法收敛性强且生

成的入侵检测规则有效而简单。

3.3.2 实验2——检测未知攻击

测试集与训练集中各类数据概率分布如表1所列，测试集中U2R和R2L类未知攻击分别占攻击总数的82.89%和62.98%。如果采用常规检测方式——将不满足4种攻击规则的数据均默认为Normal，则很多训练集中未出现的攻击将被检测为正常，造成未知攻击检测率低。

本实验采用检测规则组合Rule_attack：

IF (rule1 or rule2 or rule3 or rule4 or not(rule0))

THEN is attack

ELSE is normal

使用Rule_attack对测试集的检测性能如表5所列。由于Rule_attack只区分攻击与正常，所以得到的各类攻击的检测率为该规则检测的攻击数目与该攻击总数目之比。

Rule_attack对未知攻击saint，xterm，named的检测率分别为99.46%，92.31%，88.24%。对文献[3]中实验结果采用同样计算方式得到Kernel Miner方法(KDD'99第2名)对已知攻击和未知攻击的检测率在表5中给出。与Kernel Miner方法相比，CGREA算法生成的入侵检测规则对U2R和R2L已知攻击的检测率相对较高，且对各类未知攻击的检测率均高于Kernel Miner方法。但由于Rule_attack中规则数目与Kernel Miner方法的755个具体攻击决策树和218个类别决策树相比过少，导致误警率较Kernel Miner方法高。

表5 CGREA规则和Kernel Miner规则的检测结果(%)

攻击类型	CGREA			Kernel Miner[3]		
	已知攻击检测率	未知攻击检测率	误警率	已知攻击检测率	未知攻击检测率	误警率
Probing	99.29	78.80	0.73	99.83	74.62	0.16
DoS	99.50	15.38	2.40	99.99	14.80	0.34
U2R	71.80	21.16	0.01	51.28	16.40	3E-3
R2L	28.48	0.81	0.04	19.84	0.06	5E-3

3.3.3 性能比较

表6列出了4种机器学习方法用于入侵检测的时间和空间开销，其中CGREA算法的20条规则指规则最优个体集大小，实际检测时只选择了每类规则最优个体集的最好的1条。少量的规则和属性极大降低了检测对时间和空间的占用。

表6 与其他机器学习方法[13]性能比较

方法	训练时间/h	空间复杂度	检测能力
KDD CUP'99第1名[4]	≈24	500决策树	具体类别攻击与正常
KDD CUP'99第2名[3]	≈22	755/218决策树	具体类别/4类攻击与正常
Page_based LGP[13]	≈7.5	86条双地址格式指令	攻击与正常
CGREA	≈8.5	20条两属性规则	4类攻击与正常

结束语 本文提出的CGREA算法通过归一化预处理和文法约束来满足GEP规则的封闭性和充分性。在进化过程中采用文法生成无效GEP入侵规则个体，不断更新最优规则集来防止优良个体在遗传操作中丢失。实验结果表明：(1)生成的入侵检测规则能在保证高检测率的同时具有较低误警率，且对于未知攻击有较强检测能力；(2)在较小种群和有限代数内，不需人工干预自动获取的各类攻击检测规则只包含两个属性，提高了规则生成和入侵检测的效率。

总之，本文进行的是GEP应用于入侵检测的探索性工作，提出的CGREA算法改善了基于机器学习的入侵检测存在的未知攻击检测率不高、规则复杂或数目多、数据预处理过程复杂的问题。

参考文献

[1] Eskin E, Arnold A, Prerau M, et al. Geometric framework for unsupervised anomaly detection: detecting intrusions in unlabeled data[C]// Applications of Data Mining in Computer Security. Boston: Kluwer Academic Publishers, 2002: 77-102

[2] Bouzida Y, Cuppens F. Neural networks vs. decision trees for intrusion detection[C]// IEEE / IST Workshop on Monitoring, Attack Detection and Mitigation (MonAM). Tuebingen, Germany, Sep 2006: 81-88

[3] Levin I. KDD99 classifier learning contest LLsoft's results overview[C] // ACM SIGKDD. Boston: ACM, 2000, 1(2): 67-75

[4] Elkan C. Results of the KDD'99 Classifier Learning[C] // ACM SIGKDD. Boston: ACM, 2000, 1(2): 63-64

[5] 郑洪英，廖晓峰，倪霖，等. 进化算法及其在入侵检测中的应用[J]. 计算机科学，2007，34(11)：162-164

[6] Abraham A, Grosan C, Martin-Vide C. Evolutionary Design of Intrusion Detection Programs [J]. International Journal of Network Security (S1816-353X), 2007, 4(3): 328-339

[7] Ferreira C. Gene expression programming: mathematical modeling by an artificial intelligence(2nd edition)[M]. Berlin: Springer, 2006: 422-455

[8] 潘正君，康立山，陈毓屏. 进化计算[M]. 北京：清华大学出版社，1998：113

[9] 唐菀，杨喜敏，谢夏，等. 基于GEP的网络入侵检测规则约束及进化策略[J]. 华中科技大学学报：自然科学版，2008，36(11)：60-63

[10] Zhou C, Xiao W, Tirpak T M. Evolving accurate and compact classification rules with gene expression programming[J]. IEEE Transactions on Evolutionary Computation, 2003, 7(6): 519-531

[11] 覃俊，康立山. 多目标优化遗传算法的收敛性定义及实例研究[J]. 计算机应用与软件，2006，23(1)：1-2，22

[12] ACM Special Interest Group on Knowledge Discovery and Data Mining. SIGKDD. KDD CUP'99 DATA Set [EB/OL]. [2008-03-15]. http: // kdd. ics. uci. edu/databases/kddcup99/kddcup99. html

[13] Dong S. A Linear Genetic Programming Approach to Intrusion Detection [D]. Faculty of Computer Science, Dalhousie University, May 2003