

混合匹配规则的否定选择算法

席 亮, 张凤斌, 王大伟

(哈尔滨理工大学 计算机科学与技术学院, 黑龙江 哈尔滨 150080)

摘 要:针对目前基于免疫的检测器生成算法中,匹配规则单一造成的检测器质量低下的问题,提出一种基于混合匹配规则的否定选择算法,采用海明距离整体匹配的各段内 R 连续比特匹配的否定选择算法,即混合匹配否定选择算法,对算法的性能进行了分析.实验结果表明,该算法能够高效地产生所需要的检测器,提高了检测率,且降低了误报率.

关键词:人工免疫;入侵检测;否定选择;匹配规则;检测器

中图分类号: TP393.08 **文献标志码:** A **文章编号:** 1007-2683(2009)05-0052-04

Mix-matching Rules Negative Selection Algorithm

XI L iang, ZHANG Feng-bin, WANG Da-wei

(School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China)

Abstract: On the basis of the analysis of the deficiencies of the current DGAs (detector generating algorithm) with single affinity matching, and to solve the low quality of detectors, this paper proposes an improved DGA using mix-matching rules: hamming whole matching and R-contiguous matching in each segment, named MMNS (Mix-matching Negative Selection). And the algorithm is analyzed as well. These experiments show that, the algorithm model can efficiently generate the needed detectors, evidently enhance the detection rate and reduce the false alarm rate.

Key words: artificial immune; intrusion detection; negative selection; matching rule; detector

0 引 言

基于免疫的入侵检测系统是应用生物免疫机制,建立一个能区分自体与非自体的模型,并以检测率最大、误报率和漏报率最小为目标的模型.检测率、误报率和漏报率是衡量一个入侵检测系统性能好坏的标志.误报率是指被误检为异常的事件数目与总体事件数目的百分比,漏报率是指漏报的异常事件个数与总体事件个数的百分比.要提高入侵检测系统的性能,就要提高检测率,降低误报率和漏报

率,达到这一要求的最直接方法是生成高质量的检测器,这是入侵检测系统的核心技术.另外,无论是生成检测器阶段还是检测阶段,高效、快速、占用资源少等因素也是入侵检测系统需要考虑的性能参数.

检测器生成算法主要以否定选择算法为基础^[1],该算法的核心思想是定义一个自体集作为训练集来产生不与自体集模式匹配的检测器集,使用这些检测器来进行入侵检测.具体思想是:定义自体集为一字符串集合,随机产生一个长度相当的字符串,根据匹配规则将该字符串依次与集合中的字

收稿日期: 2008-09-18

基金项目: 国家自然科学基金项目 (60671049)

作者简介: 席 亮 (1983—),男,硕士研究生, E-mail: xljyp2002@yahoo.com.cn;

张凤斌 (1965—),男,教授,博士生导师.

串进行匹配,若匹配则重新生成随机字符串进行匹配;若该字符串不与集合中任何字符串匹配,则将其加入到检测器集中.另外,根据不同的匹配规则得到不同的检测器生成算法.

目前否定选择算法最主要的匹配规则是海明距离匹配规则^[2]和 R 连续比特匹配规则^[3-4].目前的生成算法主要是应用单一匹配规则,生成检测器的质量都不是很理想:都会生成错误的检测器,从而在检测时产生比较大的误报率;同时还会产生检测器集的黑洞问题,在检测时造成比较大的漏报率.每一种匹配规则都有自己的优点和缺点,造成以上问题的途径也是不同的.于是,本文提出了一种以否定选择算法为基础,两种匹配规则结合使用的检测器生成算法模型:以海明距离匹配规则做预处理,并将具备特定条件的备选字符串直接放入检测器集,否则进行分段 R 连续比特匹配代替整体 R 连续比特匹配再次进行筛选.经过实验表明:该算法在减少检测器总量的情况下提高了合格检测器的数量,同时也保证了算法的生成效率较高;而且在检测阶段,可以很好地降低误报率和漏报率,达到了很好的检测效果.

1 基于免疫的入侵检测系统模型

基于免疫的入侵检测系统模型关键在于区分 self/nonself(自体/非自体)来识别入侵行为.其中, self 指正常的网络行为; nonself 指异常的网络行为,即入侵行为.模型首先对网络信息进行处理,包括数据过滤、去噪、提取特定特征等,将网络信息转化成标准的格式.然后与检测器集中各个检测器进行匹配.一旦与一个检测器匹配,则认为这个网络行为是入侵行为;否则认为是正常行为并让其通过^[5].如图 1 所示.

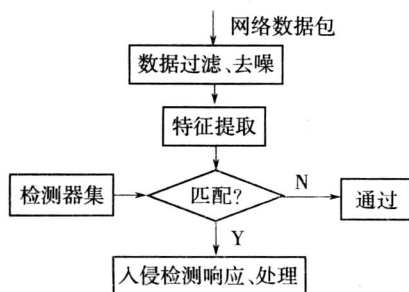


图 1 入侵检测系统模型

网络数据包都是二进制字符串,所以,检测器表示方法主要是二进制表示方法.二进制表示方法是

模拟网络数据包提取部分关键字段,提供了一个易于分析的有限空间且易于使用^[6].所以,它是目前应用最广而且最简单实用的检测器表示方法.

2 匹配规则分析与混合匹配规则算法

2.1 匹配规则分析

如前文所述,入侵检测系统的核心技术是如何生成高效合格的检测器,这对入侵检测系统来说是至关重要的.目前检测器生成算法主要是否定选择算法,基于不同的匹配规则,生成的检测器的效果也不同.现在分析一下前面提到的两种匹配规则的特点和不足.

2.1.1 海明距离匹配规则

海明距离匹配规则是最早被提出的匹配规则^[7],它主要考察的是两个数据包对应比特位匹配的字符的总数.由于该规则只是片面追求一个匹配的总数,有很明显的不足,举例如下(A为自体,B为备选检测器,下同):

A 0100 0011 1111 1111 0000
B 0100 1100 0000 0000 1111
(a)

A 0100 0011 1111 1111 0000
B 1011 1110 1001 0001 1101
(b)

(a)中两个数据段匹配了 4 位;(b)中两个数据段匹配了 5 位.根据否定选择算法将淘汰 (b)中的 B 而留下 (a)中的 B 作为检测器,但是,对于某一攻击的特征字段为字段 1 来说,显然 (b)的 B 作为检测器会更有效,而 (a)中 B 显然成了错误的检测器.可以看出,针对不同特性字段的攻击,两个数据段反映出的匹配效果可能不尽相同.如果错误地淘汰了合适的候选字符串而保留了错误的候选字符串,使其成为检测器,那么在检测器集中会产生大量的错误检测器.对于检测阶段,将造成很多误报,后果是不堪设想的.

2.1.2 R 连续比特匹配规则

R 连续比特匹配规则是在总结了海明距离匹配规则的不足后而被提出的,它主要考察的是两个数据包对应比特位上连续匹配的最大位数.该规则虽然可以比海明距离匹配规则能更好地找到两个字符串的匹配度大小,但是由于网络数据包是由特定含义的不同字段组成的,造成按照该规则计算出的亲和力并不能确切地反映出数据包间真正的相似性,

举例如下:

A 0100 0011 1111 1111 0000
B 0101 0010 1110 1110 0001
(a)

A 0100 0011 1111 1111 0000
B 1011 0011 1101 1000 1111
(b)

(a)中两个数据段 R连续匹配了 3位;(b)中两个数据段 R连续匹配了 6位.根据否定选择算法将淘汰(b)中的 B而留下了(a)中的 B作为检测器.但是可以看出,(a)中 B字段的匹配率是很高的,如果把它作为检测器,将会造成大量的误报,对于(b)中 B本来可以作为一个很好的检测器却被丢弃了,于是造成了检测器集黑洞问题,检测阶段会造成漏报问题.

2.2 混合匹配规则否定选择算法

通过以上讨论发现,单一匹配规则存在许多不足,针对这些缺点和不足,本文提出一种解决思路,描述如下:

1)摒弃了整体 R连续比特匹配方法,采用各个段分别进行 R连续比特匹配,这样避免了前面提到的整体 R连续比特匹配带来的最大匹配数较小而整体匹配度较大的弊端,降低了淘汰较好的备选字符串作为检测器的几率.但值得注意的是,如果在第一个字段不符合的情况下该备选字符串就被淘汰,而其后面的字段使其符合作为检测器的要求,从而使得它本可以作为很好的检测器保留下来却被丢弃.这样会造成检测器集黑洞问题.这个情况在海明距离匹配中进行处理.

2)在进行分段 R连续比特匹配前,先进行一次海明距离匹配,先淘汰整体匹配度较大的一些不合适的备选字符串,减轻运算负担,提高了算法的整体运算效率;同时如果某个字段无匹配位,则说明该字段可作为很好的检测值,则直接将其放到检测器集中,无需再进行分段 R连续比特匹配,这样就可很好地防止上面提到的黑洞问题,同时也提高了算法的运算效率.

综合以上思路,本文提出了一种新的检测器生成算法:先进行海明距离整体匹配,如果某个字段无匹配位,则直接将其放到检测器集中,否则再进行分段 R连续比特匹配,称之为混合匹配否定选择算法(MMNS).本算法将两种匹配规则结合在一起,但不是简单的将两种算法叠加,而是分别进行了处理和改进:首先进行海明距离匹配规则并将符合特定特

征的备选字符串直接选入检测器集,可以很好地提高算法效率以及分段 R连续比特匹配带来的黑洞问题;进行分段 R连续比特匹配可以很好地避免整体 R连续比特匹配带来的最大匹配数较小而整体匹配度较大的弊端,避免了产生错误检测器的几率.具体算法如下:

输入初始值:

H:海明距离匹配阈值

R_n : R连续比特阈值 $R_1, R_2, \dots, R_n$

N:检测器要求的数量

n:初始检测器集检测器个数(初始化为 0)

S:自体集

D:检测器集

k:字段数

BEGN

while ($n < N$)

{

随机生成字符串 s

与 S各自体进行海明距离匹配;

if(匹配值 $h \geq H$)

{

丢弃 s

continue;

}

else if(至少有一个字段无匹配位)

{

$D \leftarrow s$

continue;

}

else {

for($i = 0; i < k; i++$)

{

s 与 S各自体相应字段进行 R连续比特匹配(正则表达式法);

if($r_i < R_i$) continue;

else {

丢弃 s

break;

}

}

if($i = k$) $D \leftarrow s$

}

}

END

2.3 混合匹配算法分析

本算法是在总结了前面两种匹配规则的基础上提出的,把这两种结合在一起使用,并分别对两种匹配规则进行改进:海明距离匹配主要为了预防黑洞问题,同时提高了算法的运算效率;然后进行的分段 R 连续比特匹配是生成算法的核心,采用分段各自匹配,避免了前面提到的整体 R 连续比特匹配的缺陷,降低了生成错误检测器的几率。所以,该算法即解决了单一匹配规则带来的错误检测器问题,同时又有效地解决了单一匹配规则带来的检测器集黑洞问题,提高了生成成熟有效检测器的质量,并且在检测阶段降低了误报率,提高了报警率,达到了一个良好入侵检测系统的技术要求。

3 实验及结论

3.1 实验环境和工具

在基于 L NUX 系统下搭建的入侵检测实验平台进行测试,利用 synflooding 工具向网络中发出大量异常数据包,记录并收集数据。

3.2 实验过程

1)取 TCP 数据包的 2 个字段: IP 包头的 32 位源 IP 地址; TCP 包头的 16 位源端口号,故分别取 $R_1 = 10$, $R_2 = 6$

2)取 H 为 20。

3)自体集个数为 3 000 (事先已经生成完成)。

4)分别用海明距离匹配规则、R 连续比特匹配规则和混合匹配规则进行检测器集的生成,设定生成检测器个数 N 为 30 000;检测器生成过程如图 2 所示。

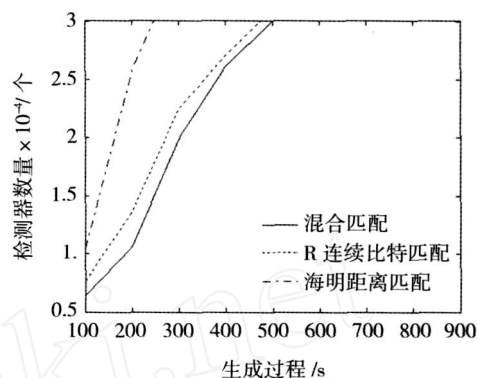


图 2 3种不同匹配规则生成的检测器过程比较

5)在这三个检测器集的检测器数量相等的情况下,用相同的数据包集对它们分别做测试,求出各自的漏报率、误报率和检测率。检测效果如表 1 所示。

6)分别取时间为 200 s 时三种匹配规则生成的检测器集来进行同样的数据测试,此时三个检测器集的检测器个数不一致。检测效果如表 2 所示。

表 1 三种不同匹配规则生成的检测器集 ($N = 30\ 000$) 检测结果

匹配规则	连接总数	正常连接数	攻击数	误报数	漏报数	漏报率 / %	误报率 / %	检测率 / %
海明距离	36 217	26 217	10 000	3 835	4 235	11.69	8.07	80.24
R 连续比特	36 217	26 217	10 000	729	6 345	17.52	2.01	80.47
混合	36 217	26 217	10 000	347	2 386	6.59	0.96	92.45

表 2 三种不同匹配规则生成的检测器集 ($t = 200\ s$) 检测结果

匹配规则	检测器个数	连接总数	正常连接数	攻击数	误报数	漏报数	漏报率 / %	误报率 / %	检测率 / %
海明距离	25 697	36 217	26 217	10 000	2 935	5 124	14.15	8.11	77.74
R 连续比特	13 660	36 217	26 217	10 000	1 739	7 382	20.38	4.80	74.82
混合	10 561	36 217	26 217	10 000	1 427	4 669	12.89	3.94	83.17

3.3 分析与结论

通过图 2 可以看出,混合匹配规则在生成检测器时的时间效率与整体 R 连续比特匹配规则基本相同,这说明首先进行海明距离整体匹配规则进行处理的效果是很明显。

从表 1 可以看出,在生成的检测器数量相同的情况下,混合匹配规则的检测效果要比前两种匹配

规则生成的检测器集的检测率高很多,且同时保证了误报率和漏报率比较低。这说明生成错误检测器的几率大大降低,同时也减少了黑洞。

同时,在保证检测效果的情况下,减少检测器数量也是研究检测器生成算法追求的目标,如表 2 所示,即使在混合匹配规则生成的检测器数量比前两
(下转第 59 页)

行.脱硫控制室设置火灾报警区域盘以监测现场的火灾情况,并应与消防及空调系统联锁;脱硫系统的火灾报警系统应与全厂火灾报警系统的选型一致,并作为全厂火灾报警系统的一个子系统,将火灾信息与全厂火灾报警的主盘进行通讯,使主盘能报警、显示火灾火警部位和控制所有监测装置.

脱硫控制系统建成后,自动化控制水平达到在现场人员的配合下,运行人员可以在控制室内以 DCS 的 LCD 及其键盘为中心,完成对整个脱硫系统的正常启、停及其运行监视、操作和故障处理等.

4 结 语

该系统通过实际安装调试后,现已正式投入运行.实际运行结果表明,该系统具有结构合理,功能完善,自动化和信息化程度高,运行稳定可靠等特点,在 Ca/S 比为 2~3 时,脱硫率可达 87% 左右,达到了国家标准《火电厂大气污染物排放标准》的要求.通过喷钙脱硫在本工程及其它 CFB 锅炉上应用表明,该喷钙脱硫工艺系统先进,设备优良,脱硫效果良好.另外该工艺投资低,占地面积小,无二次污染,运行操作简单,避免了煤粉锅炉所需的价格昂贵

的烟气脱硫装置,在技术和经济上以更为合理的方式解决了能源供应和环境保护两大难题,是循环流化床锅炉首选和前景广阔的脱硫技术装备.

参 考 文 献:

- [1] 路春美.循环流化床锅炉设备与运行[M].北京:中国电力出版社,2003:30-39.
- [2] 张焕江,钟祚群.喷钙脱硫技术在中小型循环流化床锅炉上的应用[J].电站系统工程,2008(2):53-54.
- [3] 杨吉青,周俊虎.煤粉炉喷钙脱硫技术的研究进展[J].中国电力,2005(11):97-99.
- [4] 史鸿钊,林景栋.基于模糊 PD 算法的燃烧机温控系统[J].自动化技术与应用,2008,27(7):12-13.
- [5] 王正华,周 昊.炉内喷钙脱硫对锅炉运行的影响[J].锅炉技术,2003(1):77-79.
- [6] 陈 瑞,周 征.在线自校正模糊 PD 控制器的研究[J].自动化技术与应用,2008,27(12):49-52.
- [7] 王常力,罗 安.分布式控制系统(DCS)设计与应用实例[M].北京:电子工业出版社,2004:24-26.
- [8] 聂云峰,荣丽红,高 飞,等.模糊预测控制在热电厂锅炉燃烧系统中的应用[J].自动化技术与应用,2008,27(2):19-21.
- [9] 王卫兵,李 涛,王 健.300MW 机组中压缸启动控制研究[J].哈尔滨理工大学学报,2006,11(3):110-113.

(编辑:温泽宇)

(上接第 55 页)

种方法少的情况下,检测效果相比较来说依然很不错.说明混合匹配规则算法生成的检测器质量很高,可以在较少检测器情况下依然达到检测效果的要求,同时降低了检测阶段的运算负荷.

综上所述,混合匹配规则否定选择算法生成的检测器质量要比前两种匹配规则生成的检测器质量要高很多,即生成了更多的有效检测器,而生成了较少的错误检测器,从而很好地避免了检测器集黑洞问题.而且,在保证检测器质量的同时,算法的时间代价并没有增加多少.

4 结 语

基于海明距离整体匹配的段内 R 连续比特匹配规则否定选择算法,即混合匹配规则算法,结合两种匹配规则,在达到优势互补的同时,摒弃了单一匹配规则算法各自的缺点.经过以上分析和实验表明,其生成的检测器的检测效果比前两种单一匹配规则的检测器生成算法要好很多,同时时间效率也达到

了比较理想的情况,有进行更深一步研究的价值.

参 考 文 献:

- [1] 魏 兴,张凤斌,王占锋.否定选择算法在网络入侵检测中的应用[J].哈尔滨理工大学学报,2006,11(2):90-96.
- [2] THOMAS S, JONATHAN T, CLAUDIA E. A Comparative Study of Real-valued Negative Selection to Statistical Anomaly Detection Techniques[C]. In ICARIS, 2005: 262-275.
- [3] Wenjian LUO, Xin WANG, Ying TAN. A Novel Negative Selection Algorithm with an Array of Partial Matching Lengths for Each Detector[J]. Lecture Notes in Computer Science, 2006(4193): 112-121.
- [4] 姜恩龙,张凤斌,杨 洋.基于海明匹配的 r 连续位匹配规则的研究[J].哈尔滨理工大学学报,2007,12(5):91-93.
- [5] 程永新,许家玲,陈 科.一种新型入侵检测模型及其检测器生成算法[J].电子科技大学学报,2006,35(2):235-238.
- [6] 严宣辉.免疫入侵检测模型中抗体 抗原编码的一种改进[J].计算机工程与应用,2007,43(31):151-153.
- [7] THOMAS S, JONATHAN T, CLAUDIA E. On Permutation Masks in Hamming Negative Selection[C]//In ICARIS, LNCS 4163, 2006: 122-135.

(编辑:付长缨)