应用扩张矩阵理论的攻击特征提取

陈志贤[1,2] 黄 皓[1]
(南京大学软件新技术国家重点实验室 南京 210093)[1]
(南京工业大学电子与信息工程学院 南京 210009)[2]

摘 要 近年来随着因特网的飞速发展,计算机系统也面临着越来越多的安全威胁。国内外不少研究人员为此提出了许多种基于软计算的方法用于检测网络攻击。给出了一种基于扩张矩阵理论的攻击特征提取方法,通过构造攻击子集和正常子集的扩张矩阵,建立其最优特征子集选择的整数规划模型,并利用简单遗传算法求解,最终生成可用于检测特定类型攻击的最优规则。在KDD Cup99数据集上的实验结果表明,该方法具有较高的正确检出率和可接受的低误报率。

关键词 扩张矩阵,特征子集选择,遗传算法,入侵检测
中图法分类号 TP393.08 **文献标识码** A

Attack Feature Extraction Using Extension Matrix Theory

CHEN Zhi-xian[1,2] HUANG Hao[1]
(National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China)[1]
(College of Electronics and Information Engineering, Nanjing University of Technology, Nanjing 210009, China)[2]

Abstract With the rapid development of Internet in recent years, computer systems are facing increased number of security threats. Various soft computing based approaches have been proposed to detect computer network attacks. A method for attack feature extraction based on extension matrix theory was given in this paper. By constructing extension matrix on positive and negative examples, the integer programming model for its optimal feature subset selection was built, which will be solved by simple genetic algorithm. Finally optimal rules for detection of specific attack were generated. Experimental results show the achievement of high correct detection rates and acceptable low false positive rates based on benchmark KDD Cup99 data sets.

Keywords Extension matrix, Feature subset selection, Genetic algorithm, Intrusion detection

1 引言

入侵检测是指监视着网络流量,识别出各种网络入侵,包括异常的网络行为、未经授权的网络访问以及对计算机系统的恶意攻击[1]。随着互联网技术的飞速发展,网络结构变得越来越复杂,网络安全也变得日益重要和严峻。因此,入侵检测作为一个迅速发展的新领域,已经成为网络安全研究中一个极为重要的研究方向。

入侵检测技术通常被分为两种类型:误用检测(misuse detection)和异常检测(anomaly detection)。误用检测是通过判断网络行为和预定义入侵模式是否匹配来完成检测任务;而异常检测则是通过计算网络行为和正常行为模式的偏离程度来检测入侵的,并对潜在的未知攻击发出报警。

国内外不少研究人员已经提出了许多种基于软计算的方法来检测入侵[1-6]。其理论基础涉及到模糊逻辑、人工神经网络、概率推理和遗传算法等。本文基于扩张矩阵理论的分类思想,将该算法应用于攻击特征提取。首先构造攻击子集和正常子集的扩张矩阵,然后建立该矩阵最优特征子集选择的整数规划模型并利用SGA(Simple Genetic Algorithm,简单遗传算法)求解,最后生成可用于检测特定类型网络攻击的最优检测规则。

2 扩张矩阵理论

扩张矩阵理论主要用于规则归纳学习,最早由洪家荣[7,8]提出。该算法由于具有结构清晰、理论完整、算法简练等特点,近年来得到了较大的发展,在国内外产生了显著的影响。相关概念总结如下:

设 $E=D_1\times D_2\times\cdots\times D_n$ 是 n 维有穷向量空间,其中 D_j 是有穷离散符号集,x_j 为第 j 个属性,$j\in N, N=\{1,2,\cdots,n\}$ 为变元下标集。E 中的元素 $e=\langle v_1,v_2,\cdots,v_n\rangle$ 叫做例子,其中 $v_j\in D_j$。设 PE 和 NE 是 E 的两个子集,分别叫做正例集和反例集,且 $k=|PE|, l=|NE|$。选择子是形为 $[x_j\#A_j]$ 的

到稿日期:2009-06-26 返修日期:2009-09-05 本文受国家八六三高技术研究发展计划(2007AA01Z409),国家自然科学基金项目(60673185)资助。

陈志贤(1979—),男,博士后,讲师,主要研究方向为信息安全和可信计算,E-mail:czx_leo@126.com;黄 皓(1957—),男,教授,博士生导师,主要研究方向为计算机安全与网络安全。

关系语句，其中 $A_j \subseteq D_j$，关系 $\# \in \{=, \neq, >, \geqslant, <, \leqslant\}$；公式(或者规则)是选择子的合取式，即 $\bigwedge_{j \in J}[x_j \# A_j]$，其中 $J \subseteq N$。

定义 1 已知正例 $e^+ = \langle v_1^+, v_2^+, \cdots, v_n^+ \rangle$ 及反例矩阵 NE。对于每个 $j \in N$，用“死元素” $*$ 对 v_j^+ 在 NE 中第 j 列的所有出现做代换，这样得到的矩阵叫做 e^+ 在反例集 NE 背景下的扩张矩阵，记为 $EM(e^+ | NE)$。

$$EM(e^+ | NE) = (r_{ij})_{l \times n}, r_{ij} = \begin{cases} v_{ij}^- & v_{ij}^- \neq v_j^+ \\ * & v_{ij}^- = v_j^+ \end{cases} \tag{1}$$

同样地，已知正例矩阵 PE 和反例矩阵 NE，对于每个 $j \in N$，用“死元素” $*$ 对 v_{ij}^+ 在 NE 中第 j 列的所有出现做代换，可以得到正例集 PE 在反例集 NE 背景下的扩张矩阵，记为 $EM(PE|NE)$。

$$EM(PE|NE) = (r_{ij})_{l \times n}, r_{ij} = \begin{cases} v_{ij}^- & v_{ij}^- \notin \{v_{1j}^+, \cdots, v_{kj}^+\} \\ * & v_{ij}^- \in \{v_{1j}^+, \cdots, v_{kj}^+\} \end{cases} \tag{2}$$

定义 2 在正例 e^+ 的扩张矩阵 $EM(e^+ | NE)$ 中，由分别来自不同行的 l 个非死元素 r_{ij_i}(其中 $j_i \in N$)连接组成它的一条路；$EM(e^+ | NE)$ 中的每一条路对应于一条规则 R，它可以在反例集 NE 背景下将正例 e^+ 准确识别出来。

$$R = \bigwedge_{i=1}^{l}[r_{ij_i} \# A_{j_i}] \tag{3}$$

定义 3 在扩张矩阵 $EM(PE|NE)$ 中，可能存在多条路，即存在将正例集 PE 和反例集 NE 分开的一组规则，其中包含选择子最少的规则称为最优规则。

3 最优特征子集的规划模型和求解

3.1 问题描述

一个网络连接包含有多种属性(也叫特征)，如连接的持续时间、协议类型、服务类型等。过多的属性里面往往包含了大量的冗余数据，这不但降低了机器的学习能力，而且也降低了知识模型的可理解性，从而降低了应用知识模型进行预测和决策的准确性。目前，国内外有多种特征选择的方法。而本文采用基于扩张矩阵理论的特征选取方法，通过构造正反例集的特征矩阵，求得正例集对反例集的一个覆盖，即获得正反例集的一致特征子集，那么该子集所包含的特征和正例集中该特征取值的关系式则可以形成最初的分类检测规则。这些精简的分类规则应用到入侵检测中，可以在正常网络行为的背景下将特定的网络攻击检测出来。

已知训练集 KDD Cup99，令异常连接记录集为正例集 A^+，正常连接记录集为反例集 N^-。$A^+ = \{a_1^+, a_2^+, \cdots, a_k^+\}$，$N^- = \{n_1^-, n_2^-, \cdots, n_l^-\}$，正例集和反例集的属性集合均为 $X = \{x_1, x_2, \cdots, x_{41}\}$，如 x_1 = duration，x_2 = protocol_type，x_3 = service，x_4 = flag，x_5 = src_bytes，…，可得到正例集 A^+ 在反例集 N^- 背景下的扩张矩阵：

$$EM(A^+ | N^-) = (r_{ij})_{l \times n}, r_{ij} = \begin{cases} n_{ij}^- & n_{ij}^- \notin \{a_{1j}^+, \cdots, a_{kj}^+\} \\ * & n_{ij}^- \in \{a_{1j}^+, \cdots, a_{kj}^+\} \end{cases} \tag{4}$$

则 $FS = \{x_{j_i} | r_{ij_i} \neq *, j_i \in N\}$ 是关于 A^+ 和 N^- 的一致特征子集，它可以在反例集 N^- 背景下将正例 a_i^+ $(1 \leqslant i \leqslant k)$ 准确识别出来。换句话说，特征子集 FS 里的特征和正例集 A^+ 中该特征取值的关系式构成分类检测规则，可用于区分正常的网络连接和异常的网络连接，从而达到检测入侵的目的。

3.2 基于遗传算法的特征子集选择算法

最优特征子集选择已经被证明是 NP-hard 问题。为此，很多研究工作集中于构造更好的启发式求解算法，如 FCV，AE9，HVC，GS 等，这些算法致力于在多项式时间内发现最优规则，但是所找到的不一定是最优解；或者在存在多个最优解的情况下，仅仅能够发现部分最优解[9]。

为此，基于整数规划理论建立了最优特征子集选择的整数规划模型，该模型可以更为清晰、简便地找到全部最优解，或者根据需要找出全部可行解，可以通过一次求解以获得更多的概念信息。

以扩张矩阵 $EM(A^+ | N^-)$ 为例，令 $EM(A^+ | N^-)$ 中的死元素“ $*$ ”=0，非死元素=1，建立最优特征子集的规划模型为：

$$\text{Min } Z = \sum_{j=1}^{n} x_j, s.t. \begin{cases} \sum_{j=1}^{n} r_{ij} x_j \neq 0 & 1 \leqslant i \leqslant l \\ x_j \in \{0,1\} & 1 \leqslant j \leqslant n \end{cases} \tag{5}$$

第一类约束条件要求扩张矩阵中每行必须保证有一个非死元素。第二类约束条件则表明在解空间中是否包含属性 x_j，$1 \leqslant j \leqslant n$。

遗传算法是美国密歇根大学 J. Holland 教授根据达尔文生物进化论和孟德尔遗传学说的思想提出的一种全局启发式优化算法。它利用选择、交叉和变异等遗传算子，促进解空间类似生物种群在自然界中的自然选择、优胜劣汰和不断进化，最终收敛于最优状态。遗传算法较适合于传统搜索方法所不能解决的复杂问题和非线性问题，它具有很强的全局搜索能力，并能跳过局部极值点，搜索到问题的全局最优解，已经被广泛应用于函数优化、组合优化、整数规划等领域。下面用遗传算法来求解模型(5)。

最优特征子集求解的整个流程如图 1 所示。在应用遗传算法求解问题时，3 个因素将影响到算法的效果：个体的表示、适应度函数的选择和遗传算法的参数设置。

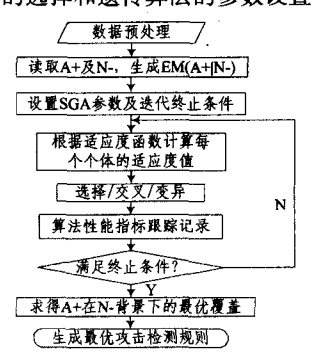

图 1 利用遗传算法求解最优特征子集的流程图

本文基因位采用二进制编码，个体可表示为 41 位的二进制字符串 $x_1 x_2 \cdots x_{41}$，其中 $x_j = 1, 0$ 表示第 j 个属性被包含或者不包含在新的个体(特征子集)中。

通常适应度函数可以直接选择为目标函数 $-\sum_{j=1}^{n} x_j$，即具有最少特征数的个体将具有最大的适应度值。式(5)第一类约束条件要求保证每一行必须有一个非死元素，否则该个体即遭淘汰，在实现中可通过给该个体加上一个极大的惩罚值，从而使其适应度值变得极小而遭淘汰。

适应度函数最终选取为：

$$f=\begin{cases}1-\frac{1}{41}\sum_{j=1}^{41}x_j & \sum_{j=1}^{41}x_j\neq 0\\ -\infty & \sum_{j=1}^{41}x_j=0\end{cases} \quad (6)$$

实验中取种群规模为 200，采用基于排序的适应度选择和最优保存策略，采用两点交叉算子，交叉概率为 0.7，变异概率为 0.0244，最大遗传代数为 180 代。

3.3 检测规则的生成

通过以上的简单遗传算法可求得 $EM(A^+|N^-)$ 的最优特征子集，然后在正例集 A^+ 中删除不属于该最优特征子集的属性，仅保留能将正反例集区分开来的属性，形成新的正例集。在此基础上过滤掉其中的重复例子，最终保留下来的每个新的例子都对应一条初始规则，它是某几个属性及其取值的合取式，将这些初始规则进行分类、合并，就得到了最优的针对特定攻击的检测规则，里面的规则都形如 $if(condition)\Rightarrow attackname$，其中 $condition$ 是最优特征子集中每个例子的属性及其取值的合取式。这些提取出来的规则或模式可以用来对当前网络连接数据进行攻击检测，也可以用来调整网络安全策略，例如修改服务器的账号管理方式等。

4 实验与分析

4.1 训练和测试数据集

MIT Lincoln 实验室的 DARPA 数据集被广泛用于评估入侵检测系统[10]。因此在实验中使用的数据集为 KDD Cup99 的网络连接数据(约 45MB，300000 多条记录)。该数据集是预分好类的，分别用数据集 2/3 和 1/3 的数据作为训练集和测试集。网络连接数据用于产生网络正常活动模式，数据记录包含的属性有连接的基本属性、内容属性和流量属性(共 41 个属性)，基本属性有连接持续时间、协议类型、服务类型、连接状态、连接的源和宿各自发出的数据长度等；内容属性包括登录失败的次数、使用 root 命令的次数、访问根的次数、创建文件的次数等；而流量属性则是通过 2 秒时间窗口计算得到的属性[11]。当然在对网络连接数据进行模型提取之前，必须对数值属性的数据进行离散化处理，以降低数据量。

4.2 实验过程及结果

实例一：获取判断攻击类型为“IPSweep”的最优规则。“IPSweep”攻击是扫描类攻击的一种，它扫描网络中哪些主机是处于监听状态，用来为其他攻击做准备的。

求解得到最优特征子集 $FS=\{x_2,x_5\}$，这表明利用 protocol_type，src_bytes 这两个属性即可以把正反例集区分开，由此得到判断攻击类型为“IPSweep”的规则：

$$protocol\_type=icmp\wedge src\_bytes\in\{8,18,1032\}\Rightarrow attack\_name=ipsweep \quad (7)$$

实例二：获取判断攻击类型为“Pod”的最优规则。“Pod”攻击是拒绝服务攻击的一种，通过不停 ping 某个主机直到其服务不可用。

求解得到最优特征子集 $FS=\{x_2,x_5,x_{23}\}$，这表明利用 protocol_type，src_bytes 以及 count 3 个属性可以把正反例集区分开，由此得到判断攻击类型为“Pod”的规则：

$$protocol\_type=icmp\wedge src\_bytes=564\Rightarrow attack\_name=pod \quad (8)$$

$$protocol\_type=icmp\wedge src\_bytes=1480\wedge count=1\Rightarrow attack\_name=pod \quad (9)$$

表 1 列出利用上述生成的入侵检测规则对训练集和测试集分别进行检测的结果。

表 1 实验结果

IPSweep attack	Training dataset	Testing dataset
True Positive rate (TP)	100%	100%
False Positive rate (FP)	0%	0%
False Negative rate (FN)	0%	0%
Pod attack	Training dataset	Testing dataset
True Positive rate (TP)	100%	100%
False Positive rate (FP)	0%	2.17%
False Negative rate (FN)	0%	0%

从表中可以看出，对于 IPSweep 攻击检测，检测规则(7)被应用于训练集和测试集时，其正确检出率均为 100%，无误报和漏报，检测效果非常理想，说明本方法生成的检测规则对 IPSweep 攻击比较敏感；对于 Pod 攻击检测，检测规则(8)、(9)应用于训练集时，其正确检出率为 100%，无误报和漏报，检测效果也非常理想；当此规则应用于测试集时，正确检出率仍为 100%，同时出现了 1 次误报，误报率为 2.17%，检测效果虽然比训练阶段略微降低，但仍然相当理想。

综上，本方法生成的检测规则在检测过程中具有较高的正确检出率，但同时会产生少量的误报，因此该特征提取方法比较适合于在一些对攻击比较敏感，但是又允许一定程度误报率的应用场合。

结束语 本文介绍了扩张矩阵理论的相关概念，并尝试将之应用于误用入侵检测中的检测规则提取。该方法具有较高的正确检出率，但是由于训练集的不完备性，它在实验中会产生少量误报。下一步的工作是尝试把本方法和人工免疫原理结合起来，以此来进一步降低漏报率，使之更适于实际的入侵检测。

参 考 文 献

[1] Gong R, Zulkernine M, Abolmaesumi P. A Software Implementation of a Genetic Algorithm Based Approach to Network Intrusion Detection[C]// Proc. of Sixth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing. Maryland: IEEE Press, 2005: 246-253

[2] El-Semary A, Edmonds J, Gonzalez-Pino J, et al. Applying Data Mining of Fuzzy Association Rules to Network Intrusion Detection[C]// Proc. of Information Assurance Workshop. West Point, NY: IEEE Press, 2006: 100-107

[3] Moradi M, Zulkernine M. A Neural Network Based System for Intrusion Detection and Classification of Attacks[C]// Proc. of the 2004 IEEE International Conference on Advances in Intelligent Systems-Theory and Applications. Luxembourg: IEEE Press, 2004

[4] Mukkamala S, Sung A, Abraham A. Modeling intrusion detection systems using linear genetic programming approach[C]// Proc. of the 17th international conference on Innovations in applied artificial intelligence. New York: Springer-Verlag, 2004: 633-642

[5] Dasgupta D, Gonzalez F. An Intelligent Decision Support System for Intrusion Detection and Response[C]// Proc. of the International Workshop on Information Assurance in Computer Networks: Methods, Models, and Architectures for Network Security. New York: Springer-Verlag, 2001: 1-14

(下转第 74 页)

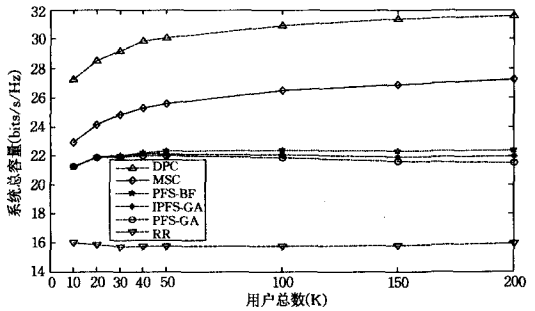

图 3　系统总速率与用户总数的关系(SNR=20dB,$n_t=4$,$n_r=2$)

在多用户 MIMO 系统下行链路中,公平的定义主要有两种,一类是基于信道资源分配的公平度,另一类是基于用户吞吐率的公平度,本文采用后一种定义,并引入公平因子来衡量算法的公平性,其定义为:$F(K)=(\sum_{k=1}^{K}x_k)^2/K\times\sum_{k=1}^{K}x_k^2$。式中 x_k 表示每个用户的平均调用比率。$F(K)$值越大,其公平性越好,其取值范围为 0～1。当 $F(K)$为 1 时,表明所用用户的调用次数基本相同,系统的公平性最好。表 1 对比了系统用户总数为 20,SNR 为 20dB 时,各种算法的公平因子。从表中可以看出,t_c 值的大小对系统的公平性影响比较小。

表 1　公平因子的比较(SNR=20dB,$K=20$)

调度算法	MSC	PFS-BF $t_c=500$	PFS-BF $t_c=100$	PFS-BF $t_c=20$	PFS-GA $t_c=100$	IPFS-GA $t_c=100$	RR
公平因子	0.4430	0.9918	0.9937	0.9950	0.9933	0.9930	1

结束语　本文针对采用块对角化预编码机制的多用户 MIMO 系统下行链路,提出了一种基于遗传算法的多用户比例公平调度算法。预编码机制可以消除多用户之间的干扰,同时为多个用户提供多数据流服务。基于穷举搜索算法的多用户调度算法复杂度太大,难以在实际系统中应用。而遗传算法能够在全局并行搜索以得到最优解或者满意解,其操作简单,适应性强。改进的遗传算法由于加入具有优秀基因的个体而加快了算法的收敛性。仿真结果显示,本文所提出的两种算法在用户非同分布时,在系统容量和公平性之间取得了良好的折中,同时算法复杂度比较低。

参 考 文 献

[1] Spencer Q H, Peel C B, Swindlehurst A L, et al. An introduction to the multi-user MIMO downlink[J]. IEEE Communication Magazine, 2004: 60-67

[2] 王德胜,朱光喜,刘应状,等. 基于虚拟 MIMO 子信道的多用户分集资源调度算法[J]. 计算机科学,2008,35(6):114-117

[3] Weingarten H, Steinberg Y, Shamai S. The Capacity Region of the Gaussian Multiple-Input Multiple-Output Broadcast Channel [J]. IEEE Transactions on Information Theory, 2006, 52(9): 3936-3964

[4] Gesbert D, Kountouris M, Heath R W Jr, et al. Shifting MIMO Paradigm: From Single User to Multiuser Communications[J]. IEEE Signal Processing Magazine, 2007, 24(5): 36-46

[5] Spencer Q H, Swindlehurst A L, Haardt M. Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels[J]. IEEE Transactions on Signal Processing, 2004, 52(2): 461-471

[6] Lai U C, Murch R D. A transmit preprocessing technique for multiuser MIMO systems using a decomposition approach[J]. IEEE Transactions on Wireless Communications, 2007, 3(1): 20-24

[7] Yoo T, Goldsmith A. Optimality of zero - forcing beamforming with multiuser diversity[C]// IEEE International Conference on Communications. 2005: 542-546

[8] Shen Z K, Chen R H, Andrews J G, et al. Sum Capacity of Multiuser MIMO Broadcast Channels with Block Diagonalization[C] //IEEE International Symposium on Information Theory. 2006: 886-890

[9] Jalali A, Padovani R, Pankaj R. Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system[C] // IEEE 51st Vehicular Technology Conference. 2000: 1854-1858

[10] Man K F, Tang K S, Kwong S. Genetic algorithms: concepts and applications[J]. IEEE Transactions on Industrial Electronics, 1996, 43(5): 519-534

[11] Holland J H. Adaptation in nature and artificial systems[D]. Massachusetts: MIT press, 1992

[12] Yoo T, Goldsmith A. Capacity and power allocation for fading MIMO channels with channel estimation error[J]. IEEE Transactions on Information Theory, 2006, 52(5): 2203-2214

(上接第 51 页)

[6] Helmer G, Wong J, Honavar V, et al. Automated discovery of concise predictive rules for intrusion detection[J]. The Journal of Systems and Software, 2002, 60(3): 165-175

[7] Hong J R. AE1: An extension matrix approximate method for the general covering problem[J]. International Journal of Computer and Information Science, 1985, 14(6): 421-437

[8] 洪家荣. 示例式学习及多功能学习系统 AE5[J]. 计算机学报, 1989, 12(2): 98-105

[9] 李敏强,寇纪淞,戴林. 示例学习和特征选择的规划模型方法[J]. 系统工程学报,2000,15(2):163-167

[10] MIT Lincoln Laboratory. 1999 DARPA Intrusion Detection Evaluation Data Set[EB/OL]. [2009-3-23]. http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/data/1999d-ata.html

[11] Lippmann R, Haines J, Fried D, et al. The 1999 DARPA Off-Line Intrusion Detection Evaluation[J]. Computer Networks: The International Journal of Computer and Telecommunications Networking, 2000, 34(4): 579-595