第43卷 第6A期
2016年6月

计 算 机 科 学
Computer Science

Vol. 43 No. 6A
June 2016

一种基于进化神经网络的混合入侵检测模型

屈洪春 王 帅

(重庆邮电大学工业物联网及网络化控制教育部重点实验室 重庆 400065)

摘 要 为了提高入侵检测系统的检测率并降低误报率,将误用检测技术和异常检测技术进行结合,以克服采用单一技术的缺陷。采用改进的进化神经网络作为检测引擎,首先,通过对遗传算法进行改进,弥补实数编码全局寻优能力差的缺陷,且降低计算的复杂度,提高进化收敛速度;然后,将改进的遗传算法和BP神经网络的LM算法进行结合,进一步克服神经网络学习阶段训练速度慢和易陷入局部最优的缺点,进而提高神经网络的分类能力和模式识别能力。采用KDDCUP99数据集作为训练与测试数据集进行实验,结果表明,基于改进的进化神经网络建立的混合入侵检测模型在数据特征规则的提取速度、检测精度以及识别新的攻击类型方面有明显改善。

关键词 入侵检测,误用检测,异常检测,遗传算法,进化神经网络

中图法分类号 TP393.08 文献标识码 A

Hybrid Intrusion Detection Model Based on Evolutionary Neural Network

QU Hong-chun WANG Shuai

(Key Laboratory of Industrial Internet of Things & Networked Control, Ministry of Education, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract In order to improve the detection rate of the intrusion detection system and reduce the false alarm rate, the misuse detection technology and anomaly detection technology were combined to overcome the single technical defect, and the improved evolutionary neural network was taken as a detection engine. Firstly, the genetic algorithm was improved to overcome the defect of the real-code poor global optimization, reduce the complexity of computation, and improve the speed of genetic algorithm evolutionary convergence. The combination of improved genetic algorithm and BP neural network LM algorithm further overcome the defects of slow training and being easy to fall into local optimum in the learning phase of neural network. Thereby, the capabilities of the neural network classification and pattern recognition increase. Using KDDCUP99 dataset as training and test data sets, experimental results show that the intrusion detection hybrid model based on evolutionary neural network can achieve significant improvement in the extraction speed of data feature rules, detection accuracy and recognizing new types of attacks.

Keywords Intrusion detection, Misuse detection, Anomaly detection, Genetic algorithm, Evolutionary neural network

1 引言

入侵检测系统(Intrusion Detection System, IDS)通过对从计算机系统或网络中关键节点处收集的信息进行分析,判断主机或网络是否遭到入侵。通常,入侵检测技术主要分为误用检测和异常检测[1]。两种检测技术各有优缺点:误用检测检测率较高,但不能检测新出现的攻击类型;异常检测能识别新出现的攻击类型,但误报率较高。另外,基于不同的数据来源,入侵检测又可分为基于网络的入侵检测[2]和基于主机的入侵检测[3]。基于主机的入侵检测主要通过对主机的审计数据和日志进行跟踪分析检测,而基于网络的入侵检测则主要通过对网络数据包的分析进行检测。

随着入侵检测技术的不断发展,智能算法也相继被运用到入侵检测中。文献[2]采用支持向量机构建了一个入侵检测系统,通过将数据映射到不同的高维空间进行分类,进而实现入侵检测。文献[4]根据免疫系统识别"自我/非自我"的能力建立一种入侵检测系统,并利用粗糙集进行特征选择优化,进一步提高系统的检测性能。文献[5]利用神经网络良好的泛化能力来识别训练数据集中未出现的攻击类型。为进一步提高检测系统的检测性能,融合多智能算法的检测系统也逐渐被提出[6-11]。基于上文的描述,本文将误用检测技术与异常检测技术进行结合,首先通过误用检测模块进行规则匹配,根据匹配程度判断网络是否遭受攻击,将未能识别的新出现的攻击类型交由异常检测模块处理,根据网络数据与正常数据轮廓的偏差程度判断该数据为攻击数据还是正常数据,通过两种技术的结合弥补了单一检测技术的缺陷。检测引擎采用BP神经网络,为了克服BP神经网络在入侵检测应用中由于其训练收敛速度慢、易陷入局部最优而导致检测系统特征

本文受中-韩美工业物联网国际联合研发中心,重庆市科技研发基地建设计划(国际科技合作)项目(cstc2013gjhz40002),重庆市基础与前沿研究计划项目(cstc2013jcyjA40014)资助。

屈洪春(1979—),男,教授,硕士生导师,主要研究方向为计算机与网络安全、智能系统与模式识别,E-mail:quhc@cqupt.edu.cn;王 帅(1989—),男,硕士生,主要研究方向为计算机与网络安全、智能系统与模式识别。

规则库建立速度慢、模式识别能力不强、检测精度不高的缺陷，本文将改进的遗传算法与 BP 神经网络进行融合，通过对遗传算法的改进，降低了计算复杂度，增强了实数编码的全局寻优能力，并根据该融合的进化神经网络提出一种基于 MGA-ANN（Modified Genetic Algorithm-Artifical Neural Network）的混合网络入侵检测模型。利用改进遗传算法良好的全局寻优能力来优化神经网络，以提高其分类精度和模式识别能力，并结合混合入侵检测技术提高入侵检测系统的性能。

2 基于 MGA-ANN 的混合网络入侵检测系统

2.1 遗传算法的改进

遗传算法常用的编码方式包括二进制编码和实数编码，二进制编码方式应用较为广泛[12]，但采用二进制编码方式处理实数域问题时，二进制与十进制映射会存在量化误差，且编码、解码计算复杂度较大；因此本文采用实数编码，实数编码大都采用简单的算术交叉算子[13]，这样会导致种群进化收敛速度慢，且全局寻优能力差，不易达到全局最优。为了克服这个缺陷，本文对遗传算法进行了相应的改进。

1）适应度及选择算子

遗传算法的适应度函数一般由目标函数变换产生，本文采用神经网络输出的均方差的倒数作为适应度。选择算子通常采用个体适应度值与总个体适应度值的比值作为该个体被选择的概率，再以轮盘赌的方式进行选择。但这样会导致在进化初期适应度值较大的个体被过多选择，容易出现早熟现象。为了弥补这个缺陷，本文采用排序选择的方法，即将个体适应度按从大到小依次排序，使个体被选择的概率随着适应度减小以公比 $(1-q)$ 的比例均匀缩小，每个个体被选择的概率由式(1)确定。

$$p_i=\beta(1-q)^{i-1} \tag{1}$$

其中：

$$\beta=\frac{q}{1-(1-q)^n} \tag{2}$$

$$q=\frac{f_{\max}}{\sum_{i=1}^{n} f_i} \tag{3}$$

p_i 为排完序第 i 个个体被选择的概率，i 为个体在序列中的序号，f_i 为第 i 个个体的适应度，$f_{\max}$ 为当前种群个体最大的适应度值，n 为种群数量。

2）交叉算子

为了提高收敛速度，本文提出一种交叉算子，使每次交叉后子代向着适应度更大的个体靠近，个体适应度较大的个体在其附近进行搜索以期得到更优的值，为每次交叉指引了导向，加快收敛速度。

当 $f(X_1)>f(X_2)$ 时，

$$X_1'=X_1+a(2\times rand-1)\times(b_{\max}-b_{\min}) \tag{4}$$

$$X_2'=X_2+rand\times(X_1-X_2) \tag{5}$$

其中，$a=2-2^{\frac{t}{T}}$，t 为当前进化的代数，T 为总进化代数，X_1，X_2 表示种群中的任意两个个体，X_1'，X_2' 分别为 X_1，X_2 交叉后的个体；$f(X_1)$，$f(X_2)$ 表示 X_1，X_2 的个体适应度，$b_{\max}$，$b_{\min}$ 为个体取值的上、下界，$rand$ 为 $(0,1)$ 范围内的随机数生成器。

当 $f(X_1)=f(X_2)$ 时，采用如下公式进行交叉操作：

$$X_1'=\alpha X_1+(1-\alpha)X_2 \tag{6}$$

$$X_2'=\alpha X_2+(1-\alpha)X_1 \tag{7}$$

其中，$0<\alpha<1$。

传统遗传算法的交叉概率一般都设定为一个固定的值，但如果设定得太大，在进化后期会导致优秀的种群个体被破坏；如果设定太小，进化初期收敛速度缓慢。使交叉概率随着种群进化代数自适应变化，对改善遗传算法性能会有很大促进作用。本文提出了如下自适应交叉概率公式：

$$P_c=P_{c_{\max}}\exp(-\frac{t}{T}) \tag{8}$$

其中，P_c 为种群交叉概率，$P_{c_{\max}}$ 为设定的最大交叉概率，本文取值 0.9；t 为当前进化代数，T 为设定的最大进化代数。由式(8)可知，交叉概率 P_c 随着种群进化代数的增大而自适应减小，既保证了进化初期种群的收敛速度，又保证了进化后期种群优秀个体不会因交叉概率过大而被破坏。

3）变异算子

为了提高遗传算法的全局搜索能力，本文采用如下变异算子：

$$X_i=b_{\min}+rand\times(b_{\max}-b_{\min}) \tag{9}$$

由上式可以看出个体 X_i 的变异范围为 $(b_{\min},b_{\max})$ 整个取值区间，保证了个体的多样性。

变异概率取值一般不宜设定得太大，取值太大将增大遗传算法搜索过程的随机性，且容易破坏种群中的优秀个体；但也不宜设定得过小，若取值过小，其会抑制早熟能力且产生的新个体的全局搜索的能力将减弱。为此，本文提出了自适应的变异概率公式：

$$P_m=P_{m_{\max}}\exp(-\frac{f_i}{f_{\max}})\frac{1}{1+\frac{t}{T}} \tag{10}$$

其中，P_m 为变异概率，$P_{m_{\max}}$ 为设定的最大变异概率，本文取值 0.3；f_i 为当前个体的适应度，$f_{\max}$ 为当前种群中个体最大的适应度；t 为当前进化代数，T 为设定的最大进化代数。由式(10)可知，适应度较小的劣质个体比适应度大的优良个体的变异概率大，且随着种群向全局最优方向的进化，变异概率也逐渐减小。

2.2 改进遗传算法与 BP 神经网络的融合

为进一步提高 BP 神经网络的检测精度和学习速度，将改进的遗传算法和 BP 神经网络进行融合，构成 MGA-ANN 融合进化神经网络。首先，用上文提出的改进的遗传算法对 BP 神经网络的权值和阈值进行全局寻优；当进化到一定代数或误差达到预设的阈值时，将种群中适应度最大的个体解析成相应的权值和阈值传给 BP 神经网络，再采用神经网络的 LM 算法进行局部寻优。遗传算法的种群规模和个体基因数由 BP 神经网络的网络结构决定。初始化种群时，由于随机初始化可能会导致种群过于集中，不利于全局寻优，收敛速度慢。为了避免这种情况，本文提出了种群相似度 S 的概念并给了相应的定义；当种群相似度 S 大于预设阈值时，就重新初始化种群，避免种群分布过于集中。

定义 1　种群中个体的平均值构成的向量为该种群的中心点 C：

$$C=\frac{\sum_{i=1}^{n} X_i}{n} \tag{11}$$

其中，X_i 为种群中任一个体，l 为个体 X_i 的长度，n 为种群的数量：

$$X_i=(X_{i,1},X_{i,2},\cdots,X_{i,l}) \tag{12}$$

定义 2　种群相似度 S 为种群中个体与中心点 C 的欧氏距离不超过与中心点 C 最远个体距离一半的个体数与种群

数量 n 的比值。

$$S=\frac{r}{n} \tag{13}$$

其中，r 为 $\frac{\|X_i-C\|<\max\|X_i-C\|}{2}$ 成立的 X_i 的个体数。

种群相似度的运用进一步提高了遗传算法的全局寻优能力；另外，初始化种群即神经网络的权值和阈值时，为了防止神经网络过早进入饱和状态，将种群初始值的范围设为 $(-1,1)$。MGA-ANN融合网络的训练步骤如下所示：

Step 1 根据 BP 神经网络的网络结构确定种群规模和个体基因数，初始化种群；

Step 2 根据式(11)—式(13)计算种群相似度 S，判断 S 是否小于阈值 k（本文 $k=0.5$，即种群数量的一半），若小于 k，则进行 Step 3，否则跳到 Step 1；

Step 3 计算个体适应度；

Step 4 根据式(1)计算个体被选择的概率，并进行选择操作；

Step 5 根据式(4)—式(8)进行交叉操作；

Step 6 根据式(9)、式(10)进行变异操作；

Step 7 计算个体适应度和总体误差，判断误差或进化次数是否满足要求，若满足要求则进行 Step 8，否则跳到 Step 4；

Step 8 将种群中个体适应度值最大的个体解析成 BP 神经网络相应的权值和阈值；

Step 9 采用 BP 神经网络的 LM 算法训练网络，进行局部寻优；

Step 10 判断神经网络输出误差或训练次数是否满足要求，若满足要求则训练过程结束，否则跳至 Step 9。

2.3 基于 MGA-ANN 的混合网络入侵检测模型

根据上文提出的融合进化神经网络建立了一种基于 MGA-ANN 的混合网络入侵检测模型，如图 1 所示。利用改进遗传算法的全局寻优能力和神经网络良好的自适应自学习能力，通过对攻击类型数据和正常类型数据的学习，分别建立网络攻击类型规则集和网络的正常行为轮廓；根据攻击类型规则集、网络的正常行为轮廓以及改进的进化神经网络良好的模式识别能力，判断网络是否遭受入侵。

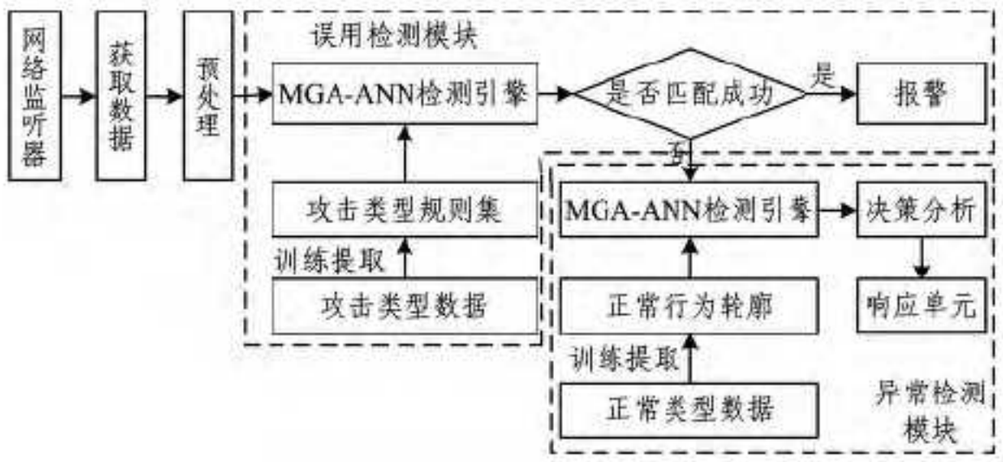

图 1 MGA-ANN 混合网络入侵检测模型

检测模型的具体工作过程如下：

Step 1 网络监听器实时监听网络的重要端口，通过抓包器获得网络数据包，并进行解析；

Step 2 通过 PCA 主元分析[14]对网络数据进行过滤，以降低计算复杂度和数据间的冗余；

Step 3 对网络数据按最大最小值方法进行归一化处理，以防止大数吃小数；

Step 4 采用上文的训练方法进行训练，用攻击类型网络数据训练误用检测引擎，构建攻击类型规则集；

Step 5 用正常类型的网络数据训练异常检测引擎，构建网络的正常行为轮廓。

Step 6 用训练好的混合检测系统实时保护网络，具体步骤为：

Step 6.1 将抓包获得的网络数据进行预处理后，输入误用检测引擎，根据模式匹配程度判断是否遭受入侵，若匹配成功，即遭到入侵，进行报警，若未匹配成功，进行 Step 6.2；

Step 6.2 将未匹配成功的数据输入异常检测引擎，若与建立的网络正常行为轮廓偏差太大，即为新出现的攻击类型，反之，即为正常类型，通过决策分析单元进行相应的响应。

3 实验分析

本文仿真测试是在 Win7 环境下进行的，用 Matlab 7.6 作为仿真平台，采用的数据集是 KDDCUP99 10% 数据集中的 20000 条数据，包括 DoS、Probe、U2R、R2L 4 大类网络攻击类型和正常类型的数据。其中 5000 条攻击类型数据作为误用检测单元训练数据集（类型分布如图 2 所示），10000 条正常类型数据作为异常检测单元训练数据集。剩下的 5000 条数据作为混合检测模型的测试数据集，包括攻击类型数据、正常类型数据以及训练数据集中没有出现的新的攻击类型数据，相应的数据类型分布如图 3 所示。BP 神经网络采用三层结构，输入层神经元的数量根据预处理后每条数据所保留属性的数量来确定。原始数据有 41 个属性，经主元分析预处理后每条数据保留 35 个属性作为神经网络的输入，确定输入层神经元数为 35。根据相关专家经验和仿真测试确定隐层神经元数为 42；误用检测引擎输出层神经元数为 4，每个神经元的输出对应一种类型，检测出某种攻击类型相应神经元输出 1，其余神经元输出 0；异常检测引擎输出神经元数为 1，输入数据与建立的网络正常行为轮廓匹配时输出 1，否则输出 0。遗传算法初始化种群规模为 50。相关仿真结果如图 4、表 1 和表 2 所示。

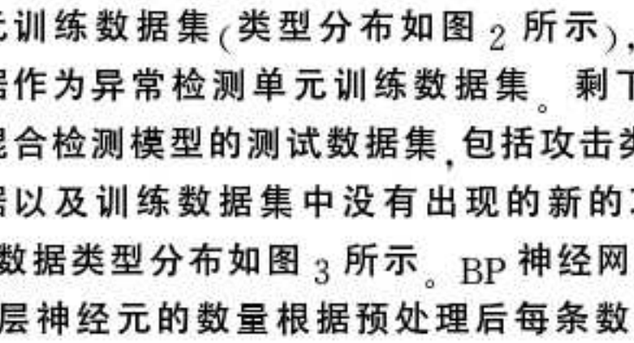

图 2 误用检测引擎训练数据类型分布图

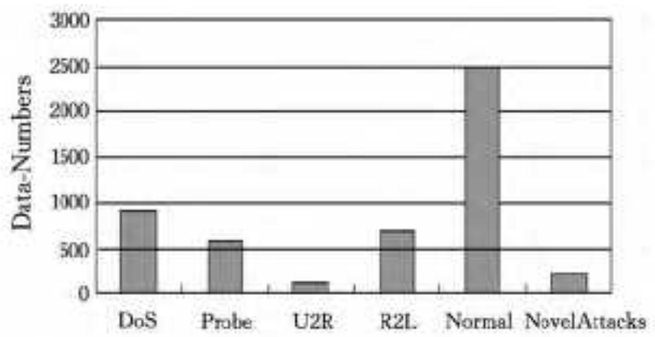

图 3 测试数据类型分布图

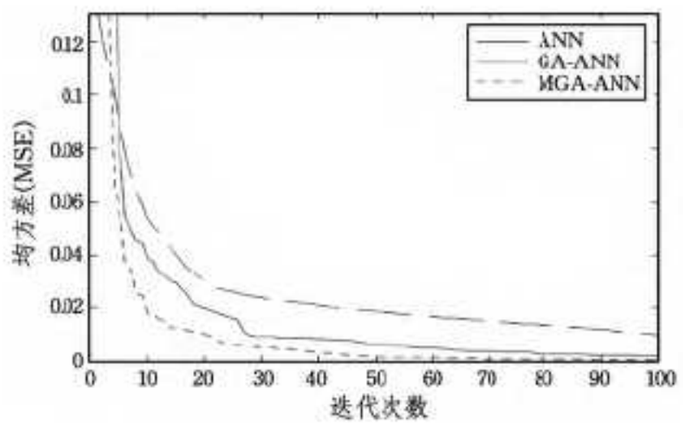

图 4 训练收敛曲线

表 1　3 种检测模型对不同攻击类型的检测率和误报率(%)

类别	GA-ANN 入侵检测系统(误用检测技术)		PSO-SVM 入侵检测系统(异常检测技术)		MGA-ANN 入侵检测系统(混合检测技术)	
	检测率	误报率	检测率	误报率	检测率	误报率
正常	87.73	4.38	88.45	5.78	96.37	1.03
DoS	85.64	3.17	84.92	5.65	90.46	1.26
Probe	78.18	4.75	78.73	6.13	88.67	2.54
U2R	78.55	5.12	77.23	6.69	85.86	2.67
R2L	76.83	4.65	77.97	5.37	81.51	2.73
Novel-Attacks	15.63	80.75	78.68	6.35	81.19	3.06

表 2　3 种检测模型的性能分析

检测模型	检测率(%)	标准差(SD)	误报率(%)	标准差(SD)	训练收敛时间(s)
GA-ANN 入侵检测系统	83.91	1.83	4.72	1.13	91
PSO-SVM 入侵检测系统	84.12	1.93	5.39	1.65	86
MGA-ANN 混合入侵检测系统	92.65	0.96	1.95	0.37	67

由图 4 可知，本文提出的 MGA-ANN 融合进化神经网络比传统 BP 神经网络和遗传神经网络(GA-ANN)在训练收敛速度及误差精度方面有显著提高，进而提高了神经网络的分类精度和模式识别的能力。通过表 1 不难看出，GA-ANN 误用检测系统对 4 种攻击类型的误报率低于 PSO-SVM 异常检测系统，但对新出现的攻击类型的检测率却非常低；而 PSO-SVM 异常检测系统对新出现的攻击类型的检测比较理想。本文提出的 MGA-ANN 混合检测系统无论对 4 种攻击类还是对新出现的攻击类型的检测率和误报率都优于误用检测系统和异常检测系统的。表 2 为 3 种检测系统总体检测性能分析，不难看出 GA-ANN 误用检测系统的总的检测率与 PSO-SVM 异常检测系统比较接近，总的误报率要稍低于后者；MGA-ANN 混合检测系统的总体检测性能要优于采用单一检测技术的 GA-ANN 入侵检测系统和 PSO-SVM 入侵检测系统，检测率和误报率的标准差也相对较小，系统鲁棒性得到了加强。

结束语　本文主要做了两方面的工作，首先，为了弥补神经网络在网络入侵检测应用中由于神经网络训练收敛速度慢、易陷入局部最优而导致基于神经网络入侵检测系统检测率不高的缺陷，对遗传算法进行改进，提高其实数编码的全局寻优能力以及进化速度；用改进的遗传算法优化 BP 神经网络，提高神经网络的训练收敛速度以及误差精度，进而达到提高神经网络的分类精度和模式识别能力的目的。其次，为了弥补误用检测系统不能识别新的攻击类型以及异常检测系统误报率高的缺陷，将误用检测技术和异常检测技术进行融合组成混合检测系统，通过仿真不难发现，基于 MGA-ANN 的混合检测系统的整体检测性能以及鲁棒性得到了改善。

参 考 文 献

[1] 阎巧，谢维信. 异常检测技术的研究与发展[J]. 西安电子科技大学学报，2002，29(1)：128-132

[2] Raju E，Sravanthi K. Network intrusion detection using Support Vector Machines[J]. International Journal of Computer Science And Management Research，2013，2(1)：1313-1319

[3] Creech G，Hu J. A Semantic Approach to Host-Based Intrusion Detection Systems Using Contiguous and Discontiguous System Call Patterns [J]. IEEE Transactions on Computers，2014，63(4)：807-819

[4] Shen J，Wang J，Ai H. An Improved Artificial Immune System-based Network Intrusion Detection by Using Rough Set[J]. Communications & Network，2012，4(1)：41-47

[5] Lee S C，Heinbuch D V. Training a neural network based intrusion detector to recognize novel attacks[J]. IEEE Transactions on Systems Man & Cybernetics Part A Systems & Humans，2001，31(4)：294-299

[6] 林冬茂，薛德黔. 一种基于无监督免疫优化分层的网络入侵检测算法[J]. 计算机科学，2013，40(3)：180-182

[7] Kim G，Lee S，Kim S. A novel hybrid intrusion detection method integrating anomaly detection with misuse detection[J]. Expert Systems with Applications，2014，41(4)：1690-1700

[8] Shirazi H M. Anomaly Intrusion Detection System Using Information Theory，K-NN and KMC Algorithms[J]. Australian Journal of Basic & AppliedSciences，2009，3(3)：2581-2597

[9] Lin S，Ying K，Lee C，et al. An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection [J]. Applied Soft Computing，2012，12(10)：3285-3290

[10] Ahmad I，Hussain M，Alghamdi A，et al. Enhancing SVM performance in intrusion detection using optimal featuresubset selection based on genetic principal components[J]. Neural Computing &Applications，2014，24(78)：1671-1682

[11] Gan X S，Duanmu J S，Wang J F，et al. Anomaly intrusion detection based on PLS feature extraction and core vector machine [J]. Knowledge-Based Systems，2013，40(1)：1-6

[12] 王丽娜，董晓梅，等. 基于进化神经网络的入侵检测方法[J]. 东北大学学报(自然科学版)，2002，23(2)：107-110

[13] 梁昔明，龙文，秦浩宇，等. 基于种群个体可行性的约束优化进化算法[J]. 控制与决策，2010，25(8)：1129-1132

[14] Han F，Liu H. High Dimensional Semiparametric Scale Invariant Principal Component Analysis[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2014，36(10)：2016-2032