基于蚂蚁算法与支持向量机的入侵检测技术*

丁 赢,殷肖川,胡 傲
(空军工程大学 电讯工程学院,陕西 西安 710077)

摘 要:基于统计检测的方法,提出了一种基于遗传蚂蚁算法与支持向量机联合优化的入侵检测技术。本算法在利用遗传蚂蚁算法对数据特征进行提取的同时,对支持向量机参数进行优化,利用遗传算法快速得到局部最优值,然后利用蚂蚁算法的全局搜索特点得到全局最优值,从而可以明显提高入侵检测正确率,缩短检测时间。仿真表明,本算法检测正确率与本文提到的其他方法相比明显提高。

关键词:遗传蚂蚁算法;支持向量机;入侵检测

中图分类号:TP393.08 **文献标识码**:A

Intrusion detection technology based on ant algorithm and support vector machine

DING Ying, YIN Xiao Chuan, HU Ao
(The Telecommunication Engineering Institute, Aire Force Engineering University, Xi'an 710077, China)

Abstract: The article presents joint optimization of intrusion detection technology based on genetic ant algorithm and support vertor machine. Exploiting genetic ant algorithm, this method extracts the data characteristics while optimizing the parameter of support vertor machine. The merit of the method is that it firstly uses genetic algorithm to obtain the local optimal value rapdily, then gets the global optima value using the global search characteristic of ant algorithm, so it can enhances the detection correctness obviously and reduces the detection time. Simulation shows that the detection correctness of the proposed method is clearly better than other methods cited in the paper.

Key words: genetic ant algorithm; support vector machine; intrusion detection

对入侵检测的研究可以追溯到20世纪80年代,入侵检测的概念由James Anderson[1]首次引入。检测技术从基于规则的匹配发展到智能检测,入侵检测系统的性能在很大程度上取决于入侵检测技术的优劣,设计有效的入侵检测技术对提高系统性能非常关键。由于入侵检测方法的重要性,已出现多种检测技术[2-4],其中的智能检测技术由于在检测正确率和自学习方面具有的良好性能而引起了广泛关注。

现有智能检测技术还存在一些不尽人意的地方。首先,现有的智能检测技术忽略了合适的特征分析技术对入侵检测的重要作用,且未考虑入侵检测所面临背景的特殊性;其次,在大规模网络环境下,现有检测技术在检测正确率、误警率、漏警率和检测效率方面不能满足要求,需设计新型的智能检测技术解决日益增长的网络规模以及处理数据量日益增大的问题。基于以上不足,本文提出一种新的解决方案,可明显提高检测正确率,且减低检测时延。

遗传算法模拟生物界物竞天择、适者生存的原则,通过交叉、变异、选择操作来进行全局优化运算;蚁群算法模拟大量无意识个体组成有意识群体的生物界行为,通过正反馈使系统趋近于有序化,从而可进行优化运算,但初期需要大量的重复运算。本文利用遗传蚂蚁算法进行特征提取,然后把支持向量机的参数作为特征加入到遗传算法染色体,从而在对入侵特征进行抽取的同时对支持向量机进行优化。

1 遗传蚂蚁算法

遗传蚂蚁算法的思想是在满足目标函数的限定条件之前采用遗传算法,充分利用遗传算法的群体性、全局收敛性、随机性、快速搜索等优势生成初始解,即产生

*基金项目:陕西省自然科学基金(项目编号:2007-24)

有关问题的初始信息素分布。随后，采用蚁群算法，在有一定初始信息素分布的情况下，最大限度地利用蚁群算法的正反馈性、并行性、求精效率高等特点求取任务调度的最优解。遗传蚂蚁算法流程如图1所示，具体步骤如下：

(1)定义目标函数和适应值函数。

(2)随机生成1组二进制或者实数编码。

(3)根据适应函数选择1对个体，并对其进行交叉计算。

(4)根据适应值函数进行变异操作，通过比较解的结果，若没生成若干组优化解，则进行选择操作，并到第(2)步重新进行交叉计算。

(5)如果生成若干优化解，则进入下一步蚂蚁算法。

(6)$nc \leftarrow 0$(nc 为迭代步数或搜索次数)，初始化各参数，即 τ_{ij} 和 $\Delta\tau_{ij}$ 的初始化，根据优化生成信息素初始分布，将 m 只蚂蚁置于 n 个节点上。

(7)将各蚂蚁的初始出发点置于当前解集中，根据蚂蚁的行进路线，计算蚂蚁 $k(k=1,2,\cdots m)$ 移动到下一节点 j 的概率 p_{ij}^{k}，根据选择概率，移动每只蚂蚁到下一节点 j，将节点 j 置于当前解集。

(8)当 m 只蚂蚁遍历 n 个节点后，最优蚂蚁圈进行信息素增加 $\Delta\tau_{ij}^{k}=Q/Z_k$，计算各蚂蚁的目标函数 Z_k $(k=1,2,\cdots m)$，记录当前的最优解。

(9)对所有路径信息素进行更新 $\tau_{ij}(t+1)=(1-\rho)\tau_{ij}(t)+\Delta\tau_{ij}(t)$，对各边弧 $\Delta\tau_{ij}\leftarrow 0$，并且 $nc \leftarrow nc+1$，若 nc 小于预定的迭代次数且无退化行为(即找到的都是相同解)，则转入(7)步重新计算，否则进入下一步。

(10)进行重新的返回迭代或输出最终结果。

2 基于条件熵的遗传蚂蚁算法参数设置

下面详细介绍基于条件熵的遗传蚂蚁算法的步骤。

2.1 遗传算法参数设置

2.1.1 遗传个体编码

遗传个体的表示涉及编码问题。借鉴参考文献[5-6]的思想，设计新的遗传个体结构。采用二进制编码与实数编码相结合的方式。遗传个体都是由定长的个体构成，表示可能的最优特征子集和对应的权重，以及SVM训练模型的参数。遗传个体的结构如图2所示，前 L 参数表示原始输入特征的个数，后 L 参数表示每个特征的权重，最后2个参数表示SVM模型中的参数C和核函数参数。这样通过优化适应度函数可以得到最优的入侵特征子集及对应的权值，以及SVM模型中的参数C及核函数参数。

选择不同的核函数构造支持向量机模型，核函数以及相应核函数参数的选择会对支持向量机的分类性能产生较大影响，进而对检测性能产生重要影响。这也是将SVM模型中的参数纳入遗传个体设计的原因，期望

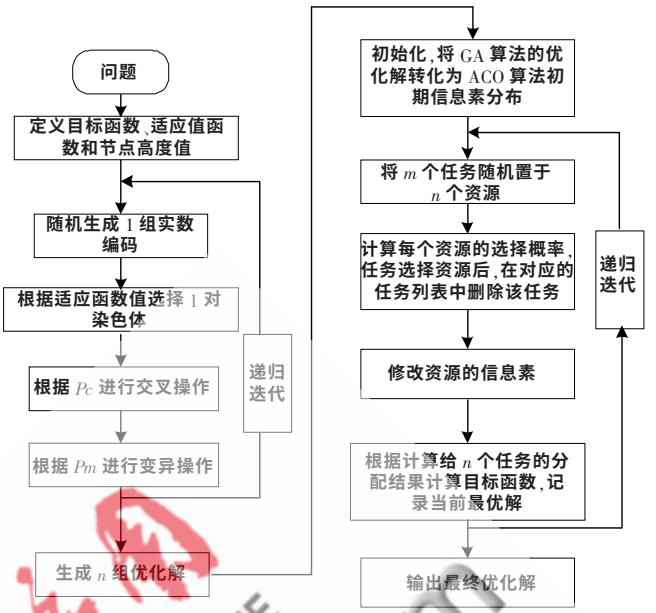

图1 遗传蚂蚁算法流程图

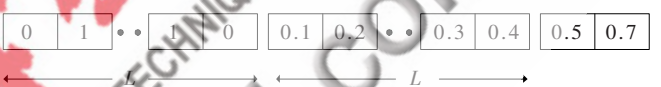

图2 遗传个体编码表示

通过遗传算法的寻优过程找到优化的SVM模型参数，以进一步提高分类性能。本文选用的核函数是径向基函数RBF(Radial Basis Function)核函数，$K(x_i,x_j)=\exp(-\varepsilon|x_i-x_j|^2)$，其中 ε 是RBF核函数参数。

2.1.2 适应度函数(目标函数)的定义

根据Helman建立入侵数学模型[7]，将系统活动的特征参数看成平稳随机过程 H 在离散时间 t_i 的取值，取值空间是有限集合 S。系统活动的状态集为 X，$X=\{$正常，异常$\}$。由于抽样的随机性，在得到特征参数后，对于系统活动状态的判断还存在风险，即某种不确定性。这种不确定性可用条件熵来度量[8]：

$$H(X|S)=\sum_{x\in X,s\in S}p(x,s)\log\frac{1}{p(x|s)} \tag{1}$$

式(1)中，$H(X|S)$ 越小，系统的不确定性也越小，越容易判断系统状态。式(1)中联合概率分布和条件概率分布根据样本数据来估计。记样本的大小为 N，$n(x,s)$ 为系统活动为 x、特征参数为 s 的样本在样本集合中出现的次数，$n(s)$ 是特征参数为 s 的样本在样本集合中出现的次数。于是有：

$$\hat{p}(x,s)=\frac{n(x,s)}{N} \tag{2}$$

$$\hat{p}(x|s)=\frac{n(x,s)}{n(s)} \tag{3}$$

以往采用条件熵进行特征抽取时，先确定选取特征数量，再采用顺序后退法[9]。然而，由于特征个数不易确定，顺序后退法剔除特征存在较大的随机性，且被剔除

的特征不能再被召回。

本文利用遗传蚂蚁算法启发式搜索的优点，以降低条件熵和分类错误率为目标，在保证高的分类正确率的同时，考虑到数据的统计特性，进一步消除冗余特征。用条件熵遗传蚂蚁算法进行特征抽取时，使用条件熵和分类错误率联合对抽取的特征进行评估。由此定义的目标函数如式(4)所示：

$$F=c_1\sum_{x\in X,s\in S}\frac{n(x,s)}{N}\log\frac{n(s)}{n(x,s)}+c_2\frac{e}{N} \tag{4}$$

式中，e 为分类错误次数，$\frac{e}{N}$ 为分类错误率，c_1、c_2 分别表示条件熵以及分类错误率在目标函数中的权重。

2.1.3 遗传算子设计

选择操作：采用轮盘赌的选择方法，同时保留种群中的最优个体直接进入下一代的进化中。

交叉操作：对实数值采用非均匀算术交叉。假设对 A、B 2 个个体进行交叉，交叉后产生的个体为：

$$A'=\alpha B+(1-\alpha)A \tag{5}$$

$$B'=\alpha A+(1-\alpha)B \tag{6}$$

变异操作：对于实数值采用均匀变异，对于二进制值则采用随机变异。

在种群演化的过程中需要保留优秀个体的模式，增强较差个体的变异能力，使算法跳出局部最优解，克服早熟的缺点。结合最小化目标函数的条件，对交叉和变异概率做以下的改进[10]：

(1)对于适应度函数值较大的个体，增加其交叉率和变异率。

(2)当种群中的大部分个体拥有相近的适应度且平均适应度与最小适应度接近时，说明此时群体可能收敛到局部最优解，此时应该增加大多数个体的交叉率和变异率，以跳出局部最优。

(3)在进化后期，交叉率应逐渐减少，以利于保留最优个体，但变异率应逐渐提高，以跳出局部最优解。

按照以上思路，提出如下交叉率 p_c 和变异率 p_m 的公式。

$$p_c=\begin{cases}\max\left[\frac{p_{c\max}+p_{c\min}}{2}+\frac{p_{c\max}-p_{c\min}}{2}F\left(\frac{f-f_{\min}}{f_{\text{avg}}-f_{\min}}\right)-\beta_1\frac{t}{T},0\right] & f\leqslant f_{\text{avg}}\\ \max\left[p_{c\max}-\beta_1\frac{t}{T},0\right] & f>f_{\text{avg}}\end{cases} \tag{7}$$

$$p_m=\begin{cases}\max\left[\frac{p_{m\max}+p_{m\min}}{2}+\frac{p_{m\max}-p_{m\min}}{2}F\left(\frac{f-f_{\min}}{f_{\text{avg}}-f_{\min}}\right)-\beta_2\frac{t}{T},0\right] & f\leqslant f_{\text{avg}}\\ \max\left[p_{m\max}-\beta_2\frac{t}{T},1\right] & f>f_{\text{avg}}\end{cases} \tag{8}$$

式中，$F(x)=\frac{e^x}{e^x+1}$，$p_{m\max}$、$p_{m\min}$ 是变异率的最大值和最小值，$p_{c\max}$、$p_{c\min}$ 是交叉率的最大值和最小值，t 是当前的进化代数，f 是当前个体的适应度，T 是进化的最大代数，β_1、β_2 是进化的代数对交叉率和变异率的影响权重，均取值为 1，$F(x)$ 是增函数。因此适应度较大个体的交叉和变异概率将高于适应度小的个体，对于适应度大于平均适应度的个体，其交叉和变异概率将大于适应度小于平均适应度的个体，由此满足改进(1)。当种群中的大部分个体拥有相近的适应度且平均适应度与最小适应度接近时，即 $f_{\text{avg}}-f_{\min}\to 0$，此时交叉率和变异率很大，满足改进(2)。当 t 逐渐增大时，交叉率减小，变异率增大，满足改进(3)。

2.2 蚂蚁算法参数设置

一般而言，蚁群算法的参数可通过反复试凑得到，但严重影响算法的运算效率。所以，国内外研究人员在蚁群算法的参数分析和优化组合方面做了大量工作。Dorigo M 等人最早对 α、β、ρ、m 等参数的选择进行了初步研究[11]；随后 BOTEE[12]等人也对下面参数的选择进行了研究。

(1)启发因子 α 的选择

启发因子 α 是表示残留信息相对重要程度的参数，其大小反映了蚂蚁在路径搜索过程中随机性因素作用的强弱，α 值越大，蚂蚁选择以前走过的路径的可能性也越大；α 值太小时，易使蚁群的搜索过程过早陷于局部最优。本算法选择 $\alpha\in[0.5,1.0]$。

(2)期望启发因子 β 的选择

β 是反映能见度相对重要程度的参数。β 的大小反映了蚂蚁在路径搜索过程中确定性因素作用的强弱，其值越大，蚂蚁选择邻近的局部最短路径的可能性也越大；β 过小，将导致蚂蚁群体陷入纯粹的随机搜索，很难找到最优解。随着 β 的增大，算法收敛速度加快。本算法选择 $\beta\in[2,2.5]$。

(3)信息素挥发度 ρ 的选择

参数 ρ 表示信息消逝程度，则 $1-\rho$ 就是信息素残留系数。ρ 的大小直接关系到蚁群算法的全局搜索能力及其搜索收敛速度，$1-\rho$ 反映蚂蚁个体相互影响的程度。本算法取 $\rho\in[0.3,0.5]$。

(4)蚁群数量 m 的选择

蚁群算法也是一种随机搜索算法，通过多个候选解组成群体的进化过程来寻求最优解。蚁群在搜索过程中表现出复杂而有序的行为，个体间的信息交流与相互协作起重要作用。本算法取 $m=25$。

3 基于遗传蚂蚁算法与支持向量机联合优化(CEGAA-SVM)的网络入侵检测算法

基于条件熵的遗传蚂蚁算法与机器学习联合优化的方法来进行网络入侵检测，CEGAA-SVM 的总体框架图如图 3 所示。

CEGAA-SVM 检测算法分为训练阶段和检测阶段两部分。训练阶段步骤如下：

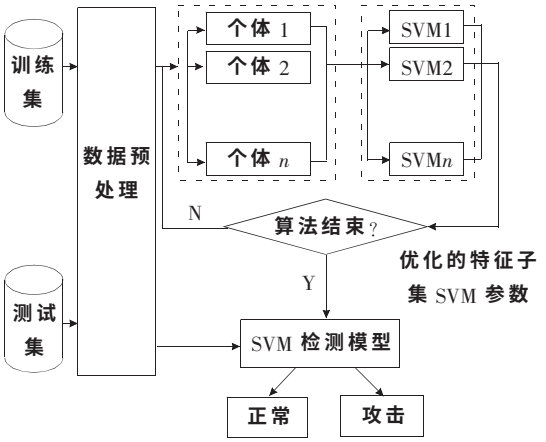

图 3 CEGAA-SVM 总体框架示意图

(1)数据预处理。设输入数据为$(d_1,d_2,\cdots,d_k)$,平均值$\bar{d}=\frac{1}{k}\sum_{i=1}^{k}d_i$,标准差$\sigma(d)=\sqrt{\frac{1}{k}\sum_{i=1}^{k}(d_i-\bar{d})^2}$,则归一化后的值$d_i'=\frac{d_i-\bar{d}}{\sigma(d)}$。

(2)随机生成初始种群。

(3)由个体的基因位确定所选择的特征、权重以及 SVM 训练模型参数,根据适应度函数计算每个个体的适应度函数值,计算交叉率和变异率。

(4)对被选中的 2 个个体进行交叉操作,产生后代个体。对被选中的个体进行变异操作。根据轮盘赌选择法按照个体的适应度函数值大小对个体进行选择操作,保留最优个体直接进入下一代种群。

(5)重复执行步骤(3),直到进化到最大代数,得到 1 个可行解的子集,设子集中解的个数为 n。

(6)根据上一小节,设置一群算法的初始参数,并以遗传算法的输出作为信息素的初始值。

(7)对于蚂蚁 k,计算其移动到下一节点 j 的概率p_{ij}^{k},根据选择概率移动每只蚂蚁到下一节点 j,将节点 j 置于当前解集。

(8)当 m 只蚂蚁遍历 n 个节点后,最优蚂蚁圈进行信息素增加$\Delta\tau_{ij}^{k}=Q/Z_k$,计算各蚂蚁的目标函数 $Z_k(k=1,2,\cdots m)$,记录当前的最优解。

(9)对所有路径信息素进行更新$\tau_{ij}(t+1)=(1-\rho)\tau_{ij}(t)+\Delta\tau_{ij}(t)$,对各边弧$\Delta\tau_{ij}\leftarrow 0$,并且$nc\leftarrow nc+1$,若 nc 小于预定的迭代次数且无退化行为,则转入第(7)步重新计算,否则进入下一步。

(10)进行重新的返回迭代或输出最终结果。

检测阶段根据选择的最优特征子集及其权重和 SVM 优化参数建立 SVM 检测模型,对待分类个体进行判断。

4 仿真与分析

4.1 仿真数据集

实验数据来源于 KDDCUP99 数据集,分为训练集和测试集,该数据集由麻省理工学院林肯实验室提供。每条连接信息包含 41 维特征,包括基本特征集、内容特征集、流量特征集和主机流量特征集。每个网络连接都被标记为正常或攻击,取值包括 Normal、Probe、DOS、U2R 和 RZL 5 种类型。

在仿真中,将实验数据集分为 4 个部分:Probe、DOS、R2L 和 U2R 数据集,各数据集样本数量如表 1 所示,测试集从 KDDCUP99 数据集的测试集中随机抽取。

表 1 仿真数据集的样本数量

数据采集	训练集		测试集	
	正常样本	攻击样本	正常样本	攻击样本
Probe	500	300	500	300
Dos	500	300	500	300
R2L	500	300	500	300
U2R	50	30	50	30

CEGAA 的参数设置如表 2 所示。

表 2 CEGAA 的参数设置表

参数名		参数值
遗传算法	种群大小	100
	个体维数	84
	进化最大代数	500
	交叉率	{0.4,0.8}
	变异率	{0.08,0.15}
	C_1,C_2	0.5
Fengxue 蚁群算法	启发因子 α	$\alpha\in[0.5,1.0]$
	期望启发因子 β	$\beta\in[2,2.5]$
	信息素挥发度 ρ	$\rho\in[0.3,0.5]$
	蚁群数量 Nc	25
	迭代次数	200

4.2 仿真结果及分析

实验在 Matlab7.0 环境中运行。CEGAA-SVM 的实验结果如表 3 所示。通过对各类数据集(Probe、DOS、U2R、R2L)的测试集进行实验,由仿真结果可以看出,对特征进行维数约减和空间变换后,不仅入侵特征的数量基本减少了一半,而且正确检测率仍然取得了满意的结果。

表 3 CEGAA-SVM 试验结果

数据集	特征数	SVM 模型参数	正确检测率/(%)
Probe	16	C=98.4,ε=2.15	99.2
Dos	18	C=304.7,ε=0.14	98.9
R2L	18	C=315.2,ε=0.45	99.7
U2R	21	C=552.1,ε=0.14	98.3

为进一步验证 CEGAA-SVM 的性能,分别采用标准 SVM 进行入侵检测、采用标准遗传算法进行特征抽取,以分类错误率为适应度函数,结果如表 4 所示。

根据对特征数目、正确检测率和检测时间的比较可以看出,采用特征抽取与 SVM 联合优化的 GA-SVM 和 CEGAA-SVM,在正确检测率和检测时间上明显优于传

表 4 CEGAA-SVM 与 SVM、GA-SVM 性能比较

数据集	算法	特征数目	正确检测率/(%)	检测时间/ms
Probe	S-SVM	41	85	0.485
	GA-SVM	24	96.2	0.283
	CEGAA-SVM	16	99.2	0.260
Dos	S-SVM	41	83.3	0.484
	GA-SVM	20	96.2	0.273
	CEGAA-SVM	18	98.9	0.264
R2L	S-SVM	41	87.4	0.024
	GA-SVM	21	94.3	0.017
	CEGAA-SVM	18	99.7	0.012
U2R	S-SVM	41	87.3	0.482
	GA-SVM	24	94.6	0.276
	CEGAA-SVM	21	98.3	0.265

统 SVM 算法。而采用条件熵遗传算法的 CEGAA-SVM 不仅降低了特征的维数，而且检测率比 GA-SVM 高，以 Probe 数据集为例，正确率提高了 3.0%。在正确率提高的同时也降低了检测时延，检测时延减少了 0.023 ms。表 5~表 8 是采用 CEGAA-SVM 分别对 4 个数据集检测结果的混淆矩阵。

表 5 Probe 数据集混淆矩阵

	Normal	Probe	误检率/(%)	漏检率/(%)
Normal	490	10	2.0	
Probe	3	297		1.0

表 6 Dos 数据集混淆矩阵

	Normal	Dos	误检率/(%)	漏检率/(%)
Normal	487	13	2.6	
Dos	0	300		0

表 7 U2R 数据集混淆矩阵

	Normal	U2R	误检率/(%)	漏检率/(%)
Normal	49	1	2.0	
U2R	0	30		0

表 8 R2L 数据集混淆矩阵

	Normal	R2L	误检率/(%)	漏检率/(%)
Normal	485	13	3.0	
R2L	2	298		0.6

从分类正确率的角度来看，本文所提供算法的平均性能要优于其他文献的算法，特别是在较难检测的 U2R 数据集上取得了满意的检测效果。

通过以上的仿真结果，可以得到如下结论：

(1)将特征分析技术和分类算法相结合的入侵检测技术能够有效地提高检测精度和检测效率。

(2)将特征抽取和分类模型进行联合优化的入侵检测技术，特征选择时考虑到数据的统计特性，在得到优化的特征向量的同时也得到与之相匹配的分类检测模型。算法不论是在检测正确率还是在检测时延方面都比较理想，对检测正确率不高的 U2R 数据集性能提升非常明显，算法的漏警率和误警率也较低。

(3)使用权重表示各个特征的重要程度，对特征空间进行线性变换。这种方法相对于简单地选择或丢弃某些特征能够取得更好的检测效果。

本文提出结合条件熵遗传蚂蚁算法和支持向量机的入侵检测技术 CEGAA-SVM，将条件熵遗传蚂蚁算法用于最优特征子集的选择，同时进行 SVM 模型参数优化，寻找最佳特征子集和与之相匹配的 SVM 检测模型，从而实现对入侵的检测。实验表明，采用 CEGAA-SVM 所提取的特征数量为 SVM 算法的一半，为 GA-SVM 算法的 83%左右。CEGAA-SVM 的正确率也有所提高，比 SVM 和 GA-SVM 算法分别提高 10%和 3%，其漏检率和误检率也有明显下降。可见，CEGAA-SVM 是一种有效的入侵检测算法。

参考文献

[1] DAKE R. BIOMETRIC security [J]. Dr. Dobb's Journal, 2001,26(11):93-96.
[2] HONG J. Plan reeognition through goal graph analysis[C]. Proeeedings of 14th ECAI. Berlin, Gennany, 2000:496-500.
[3] 刘日仙，谷文祥，殷明浩.智能规划识别及其应用的研究[J].计算机工程，2005，31(15):169-171.
[4] 段新生.证据理论与决策[M].北京:中国人民大学出版社，1993.
[5] WHITE G B, FISCH E A, POOCH U W. Cooperating security manager: a peer-based intrusion detection system[J]. IEEE Network, 1996,10(1):20-23.
[6] ASAKA M, TAGUCHI A, GOTO S. The implementation of IDA: an intrusion detection agent system. Proceedings of the FIRST Conf[C]. Brishane,1999.
[7] 陈志文，王开云，姜建国.网络入侵检测系统的警报合成算法设计[J].信息与电子工程，2005，3(3):182-185.
[8] VALEU F, CUI Y. An intrusion alert correlator based on prerequisites of intrusion. Technical Report, TR-2002-01, Department of Computer Science, North Carolina State University, 2002.
[9] 张国宣，孔锐，施泽生，等.基于核聚类方法的多层次支持向量机分类树[J].计算机工程，2005，31(5).
[10] MURPHY S. The advanced encryption standard (AES)[J]. Information Security Technical Report, 1999,4(4):12-17.
[11] SEHNEIER B. Cryptographic design vulnerabilities[J]. IEEE Computer.1999,31(9):29-33.
[12] NACEACHE D. Padding attacks on RSA [J]. Information Security Technical Report, 1999,4(4):28-33.

(收稿日期：2009-10-12)

作者简介：
丁赢，男，1983 年生，硕士研究生，主要研究方向：网络信息安全。