

# 一种高效的面向轻量级入侵检测系统的特征选择算法

陈 友<sup>1),2)</sup> 沈华伟<sup>1),2)</sup> 李 洋<sup>1),2)</sup> 程学旗<sup>1)</sup>

<sup>1)</sup>(中国科学院计算技术研究所 北京 100080)

<sup>2)</sup>(中国科学院研究生院 北京 100039)

**摘 要** 特征选择是网络安全、模式识别、数据挖掘等领域的重要问题之一. 针对高维数据对象, 特征选择一方面可以提高分类精度和效率, 另一方面可以找出富含信息的特征子集. 文中提出一种 wrapper 型的特征选择算法来构建轻量级入侵检测系统. 该算法采用遗传算法和禁忌搜索相混合的搜索策略对特征子集空间进行随机搜索, 然后利用提供的数据在无约束优化线性支持向量机上的平均分类正确率作为特征子集的评价标准来获取最优特征子集. 文中按照 DOS, PROBE, R2L, U2R 4 个类别对 KDD1999 数据集进行分类, 并且在每一类上进行了大量的实验. 实验结果表明, 对每一类攻击文中提出的特征选择算法不仅可以加快特征选择的速度, 而且基于该算法构建的入侵检测系统在建模时间、检测时间、检测已知攻击、检测未知攻击上, 与没有运用特征选择的入侵检测系统相比具有更好的性能.

**关键词** 特征选择; 遗传算法; 禁忌搜索; 线性支持向量机; 入侵检测系统

**中图法分类号** TP309

## An Efficient Feature Selection Algorithm Toward Building Lightweight Intrusion Detection System

CHEN You<sup>1),2)</sup> SHEN Hua-Wei<sup>1),2)</sup> LI Yang<sup>1),2)</sup> CHENG Xue-Qi<sup>1)</sup>

<sup>1)</sup>(Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080)

<sup>2)</sup>(Graduate University of Chinese Academy of Sciences, Beijing 100039)

**Abstract** Feature selection is one of the most important problems in network security, pattern recognition and data mining areas. For high dimension data, feature selection not only can improve the accuracy and efficiency of classification, but also discover informative subset. This paper proposes a new feature selection algorithm aiming at building lightweight intrusion detection system (IDS) by (1) using a hybrid strategy of genetic algorithm and tabu search (GATS) as search strategy to specify a candidate subset for evaluation; (2) using modified linear Support Vector Machines (SVMs) iterative procedure as wrapper approach to obtain the optimum feature subset. The authors have examined the feasibility of the feature selection algorithm by conducting several experiments on KDD1999 intrusion detection dataset which was categorized as DOS, PROBE, R2L and U2R. The experimental results show that the approach is able not only to speed up the process of selecting important features but also to guarantee high detection rates. Furthermore, the experiments indicate that intrusion detection system with a combination of feature selection algorithm has better performances than that without feature selection algorithm in terms of building time, testing time and detection rates.

**Keywords** feature selection; genetic algorithm; tabu search; linear support vector machines; intrusion detection system

收稿日期: 2007-03-05; 修改稿收到日期: 2007-05-22. 本课题得到国家“九七三”重点基础研究发展规划项目基金(2004CB318109)和国家信息安全计划项目基金(2005C39)资助. 陈 友, 男, 1981 年生, 博士研究生, 主要研究方向为网络安全、数据挖掘. E-mail: chenyou@software.ict.ac.cn. 沈华伟, 男, 1982 年生, 博士研究生, 研究方向为复杂网络、信息安全. 李 洋, 男, 1978 年生, 博士研究生, 研究方向为计算机网络信息安全、基于数据挖掘和机器学习方法的入侵检测技术等. 程学旗, 男, 1971 年生, 博士, 研究员, 研究领域为 Internet 高性能软件、智能信息处理、信息安全等.

## 1 引言

入侵检测是一种通过收集和分析被保护系统信息,从而发现入侵的技术.它的主要功能是对网络和计算机系统进行实时监控,发现和识别系统中的入侵行为或企图,给出入侵警报.可将入侵检测看作是区别系统状态是“正常”还是“异常”的二分类问题<sup>[1]</sup>.对入侵检测系统的要求首先是正确性,其次是实时性.只有检测速度快,才能及时处理网络中传输的海量数据,不会因为速度慢而丢失信息,造成漏报,更能及时采取措施,将入侵带来的损失降到最低.随着网络速度的高速提升,入侵检测系统面临的一个主要问题是检测速度低,计算资源耗用大,来不及实时处理网络中传输的海量数据,并且这个问题变得越来越严重.检测速度已经成为入侵检测系统实时性要求的一个重要指标,如何保证在正确检测的前提下,开发出检测速度快的轻量级入侵检测系统已成为当前研究的热点.因提取和处理的特征数目过多是导致入侵检测系统速度下降的主要原因,很多研究者通过特征选择来解决这个问题.特征数目和分类器性能之间并不存在线性关系,当特征数量超过一定限度时,会导致分类器性能变坏.实际上,有些特征没有包含或者包含极少的系统状态信息,它们对检测结果几乎没有影响.所以特征选择——去除冗余特征,保留能够反映系统状态的重要特征是提高检测速度的一个有效方法.一般情况下,只有特征向量中包含足够的类别信息,才能通过分类器实现正确分类,而特征中是否包含足够的类别信息却很难确定.为了提高分类器识别率总是最大限度地提取特征信息,结果不仅使特征空间维数增大,而且特征之间可能存在较大的相关性和冗余性.这给基于这些特征建立的分类器的实现带来了很大的困难,因而需要在不降低分类精度的前提下,尽量降低特征空间的维数,这也是本文提出特征选择算法的目的.

特征选择有 filter 和 wrapper 两种模型<sup>[2]</sup>,filter 模型利用数据本身的特性作为特征子集的度量指标,而 wrapper 模型利用机器学习算法的分类正确率作为特征子集的度量指标.一般来说 filter 模型的效率高,效果差;wrapper 模型的效率低,效果好.为了解决两种特征选择模型存在的问题,发挥它们的优势,很多学者提出了结合 filter 模型和 wrapper 模型的 hybrid 模型<sup>[3]</sup>.虽然 hybrid 模型在性能上有一定的提高,但是效果不理想.本文采用 wrapper 模

型设计一种高效的特征选择算法,它不仅可以克服 wrapper 模型计算资源耗用大的缺点,而且选择出的特征很大程度上提高了入侵检测系统的检测率.

## 2 相关工作

很多研究者们通过提出高效率的分类器来建立轻量级入侵检测系统,但这项工作进展缓慢,并且效果不明显.因此本文从另一个角度提出一种高效的特征选择算法来构建轻量级入侵检测系统.轻量级入侵检测系统主要包括特征选择和分类器,而特征选择包括搜索策略和评价标准两部分.在搜索策略方面,针对大数据集,文献[4]中 Jain 等人提出了诸如正向搜索、反向搜索、顺序搜索等启发式搜索策略,文献[5]中 Kudo 等人提出了比启发式搜索更有优势的随机搜索策略,如遗传算法.在大规模数据集上的特征选择,这些搜索策略的计算资源耗用大,收敛速度慢,并且在很多情况下得到的是局部最优解.针对上述缺点,本文提出一种遗传算法(Genetic Algorithm, GA)<sup>[6]</sup>和禁忌搜索(Tabu Search, TS)<sup>[7]</sup>相结合的混合搜索策略 GATS, GATS 兼具了 GA 与 TS 的优点.在评价标准方面,支持向量机以其出色的学习性能,已经成为继神经网络之后新的研究热点.它<sup>[8]</sup>建立在统计学习理论基础之上,能够很好地解决高维数、非线性和局部最小性等实际问题.文献[9]中 Weston 介绍了一种基于支持向量机的特征选择算法,依据该算法可以选出那些分类信息明确的特征.文献[10]中 Grandvalet 介绍了一种可以自动计算属性间相互关系的算法.文献[11]中 Guyon 使用递归属性排除及线性向量机作属性评估.这些算法表明支持向量机在特征选择领域的应用,但是这些算法在特征子集的评估速度上有待提高.本文根据统计学习理论和最优化理论建立了线性支持向量机的无约束优化模型,并且对此模型给出了一种有效的近似解法——极大熵方法<sup>[12-13]</sup>,为快速求解支持向量机优化问题提供了一种新途径.

在入侵检测系统分类器上,支持向量机的应用也很广泛.文献[14]中 Lee 提出了 RSVM(Reduced Support Vector Machines),它首先从训练样本中选择一个子集作为支持向量,并且解决一个更小的 QP 问题.文献[15-17]中 Kim 等在支持向量机核函数参数上进行了一定的优化.本文的重点不是分类器,而是为建立分类器模型服务的特征选择算法.文献[18]详细介绍了当前典型的特征选择算法,本文提出的特征选择算法要优于文献[18]这些典型的特

征选择算法,详细的比较将在实验部分给出.本文使用的分类器是在特征选择的过程中优化产生的.

3 特征选择算法

本文提出的特征选择算法包括搜索策略——遗传算法<sup>[6]</sup> (Genetic Search) 和禁忌搜索<sup>[7]</sup> (Tabu Search) 相结合的混合搜索策略 GATS, 评估标准——无约束优化线性支持向量机模型.

3.1 无约束优化线性支持向量机模型

支持向量机<sup>[8]</sup> (Support Vector Machine, SVM) 是 Vapnik 等人提出的一类新型机器学习方法. 它建立在统计学习理论基础之上, 能够较好地解决小样本、高维数、非线性和局部最小点等实际问题. 其思想由图 1 说明, 在图中, 实心点和空心点代表两类样本,  $H$  为分类超平面,  $H_1, H_2$  分别代表各类中离  $H$  最近的样本且平行于  $H$  的面, 它们之间的距离称为分割距离. 所谓最优分类面就是要求不但能将两类正确分开, 而且使分类间隔最大.  $H_1, H_2$  上的样本点称为支持向量.

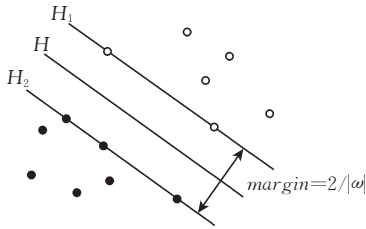


图 1 线性可分情况下的最优分类面

SVM 是建立在结构风险最小化原则基础上的. 对于训练样本为  $\{(\mathbf{x}_i, y_i)\}_{i=1}^l \subset \mathcal{R}^n \times \{1, -1\}$  的二分类问题, 根据统计学理论, 可建立如下标准的线性 SVM 模型 A

$$\begin{cases} \min & \frac{1}{2}(\boldsymbol{\omega}^T \boldsymbol{\omega}) + C \sum_{i=1}^l \xi_i \\ \text{s. t.} & y_i((\boldsymbol{\omega}^T \mathbf{x}_i) + b) \geq 1 - \xi_i, \quad i=1, 2, \dots, l \\ & \xi_i \geq 0, \quad i=1, 2, \dots, l \end{cases} \quad (1)$$

其中,  $C > 0$  为正则化参数,  $\xi_i (i=1, 2, \dots, l)$  为松弛变量,  $\boldsymbol{\omega} \in \mathcal{R}^n$  为分类超平面的法向量,  $b \in \mathcal{R}$  为阈值. 利用优化理论中的 KKT 条件和对偶理论, 可得对偶优化模型 A'

$$\begin{cases} \max & \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^T \mathbf{x}_j) \\ \text{s. t.} & \sum_{i=1}^l y_i \alpha_i = 0 \\ & 0 \leq \alpha_i \leq C, \quad i=1, 2, \dots, l \end{cases} \quad (2)$$

其中,  $\alpha_i (i=1, 2, \dots, l)$  为 Lagrange 乘子. 优化问题 A' 是一个凸二次规划问题, 其局部最优解即为全局最优解. 若  $\boldsymbol{\alpha}^* = (\alpha_1^*, \alpha_2^*, \dots, \alpha_l^*)^T$  为模型 A' 的最优解, 则

$$\boldsymbol{\omega}^* = \sum_{i=1}^l \alpha_i^* y_i \mathbf{x}_i \quad (3)$$

根据 KKT 互补条件, 最优解必满足

$$\alpha_i (y_i (\boldsymbol{\omega}^T \mathbf{x}_i + b) - 1 + \xi_i) = 0, \quad i=1, 2, \dots, l \quad (4)$$

$$(C - \alpha_i) \xi_i = 0, \quad i=1, 2, \dots, l \quad (5)$$

由式(3)~(5)可知, 对应于 Lagrange 乘子  $\alpha_i = 0$  的样本对分类问题不起什么作用, 而只有对应于 Lagrange 乘子  $\alpha_i > 0$  的样本(支持向量)对计算  $\boldsymbol{\omega}^*$  起作用, 从而决定分类结果. 支持向量通常只是全体样本中的很少一部分. 求解上述问题后得到的广义最优线性分类器是

$$f(\mathbf{x}) = \text{sgn}\{(\boldsymbol{\omega}^*{}^T \mathbf{x}) + b^*\} = \text{sgn}\left\{\sum_{i=1}^l \alpha_i^* y_i (\mathbf{x}_i^T \mathbf{x}) + b^*\right\} \quad (6)$$

其中  $\text{sgn}(\cdot)$  为符号函数,  $b^*$  为分类的阈值, 可通过任意一个支持向量求得.

模型 A 与 A' 都是约束条件下的优化模型, 为了得到线性支持向量机的无约束优化模型, 不妨定义

$$g_i(\boldsymbol{\omega}, b) \triangleq 1 - y_i(\boldsymbol{\omega}^T \mathbf{x}_i + b), \quad i=1, 2, \dots, l \quad (7)$$

则由式(4), (5)可得

$$\xi_i = \max\{0, g_i(\boldsymbol{\omega}, b)\}, \quad i=1, 2, \dots, l \quad (8)$$

将其代入标准的线性 SVM 模型 A 中, 可得线性 SVM 的无约束优化模型 B

$$\min \quad \frac{1}{2} \boldsymbol{\omega}^T \boldsymbol{\omega} + C \sum_{i=1}^l \max\{0, g_i(\boldsymbol{\omega}, b)\} \quad (9)$$

由最优化理论知, 模型 A 与模型 B 等价, 模型 B 目标函数中的前两项恰好体现了统计学习理论中的结构风险最小化原则. 其中前一项反映了模型的置信范围, 后一项反映了模型的训练误差. 注意到

$$\|\boldsymbol{\xi}\|_1 = \sum_{i=1}^l \max(0, g_i(\boldsymbol{\omega}, b)) \quad (10)$$

$$\|\boldsymbol{\xi}\|_\infty = \max\{0, g_1(\boldsymbol{\omega}, b), \dots, g_l(\boldsymbol{\omega}, b)\} \quad (11)$$

如果用向量  $\boldsymbol{\xi} = (\xi_1, \xi_2, \dots, \xi_l)^T$  的  $\infty$  范数来度量模型的训练误差, 则可得无约束优化模型 C

$$\begin{aligned} \min \Phi(\boldsymbol{\omega}, b) = \\ \frac{1}{2} \boldsymbol{\omega}^T \boldsymbol{\omega} + C \max\{0, g_1(\boldsymbol{\omega}, b), \dots, g_l(\boldsymbol{\omega}, b)\} \end{aligned} \quad (12)$$

与模型 A, A' 相比, SVM 无约束优化模型 B, C 在数学形式上更加简洁明了. 但由于优化问题 B, C 是一个和最大值函数有关的一类不可微优化问题, 从而

给求解带来困难。一条可行的途径是利用光滑化技术将不可微优化问题转化为可微优化问题，从而易于求解。本文利用极大熵方法作为求解优化问题 C 的一种近似解法。极大熵方法的基本思想是<sup>[12-13]</sup>：对于极大极小问题  $\min \phi(\mathbf{x}) = \max \{f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_m(\mathbf{x})\}$ ，利用最大熵原理推导出一个可微函数  $\phi_p(\mathbf{x} = \frac{1}{p} \ln(\sum_{i=1}^m \exp(p f_i(\mathbf{x})))$ ，通常称为极大熵函数。用该可微函数来逼近最大值函数  $\phi(\mathbf{x})$ ，从而把不可微优化问题转化为可微优化问题，使问题简化。通过引入极大熵函数，问题 C 的求解被转化为如下的无约束优化问题 D

$$\min \Phi_p(\omega, b) = \frac{1}{2} \omega^T \omega + \frac{C}{p} \ln(1 + \sum_{i=1}^l \exp(p g_i(\omega, b))) \quad (13)$$

其中,  $p > 0$  是参数。由极大熵函数的逼近性质<sup>[19]</sup>，可以得到如下定理<sup>[20-21]</sup>。

**定理 1.** 当  $p \rightarrow \infty$  时,  $\Phi_p(\omega, b)$  一致收敛于  $\Phi(\omega, b)$ ，并且

$$\Phi(\omega, b) \leq \Phi_p(\omega, b) \leq \Phi(\omega, b) + \frac{C}{p} \ln(l+1).$$

**定理 2.** 对任意给定的  $p > 0$ ,  $\Phi_p(\omega, b)$  在整个  $\mathcal{R}^{n+1}$  空间上是凸函数。

由定理 1 可以看出，当参数  $p$  趋于无穷大时， $\Phi_p(\omega, b)$  在整个  $\mathcal{R}^{n+1}$  空间上一致逼近  $\Phi(\omega, b)$ ，因此问题 C 和 D 是等价的。实际计算时， $p$  可取一比较大的限值，只用一次无约束优化计算，就可得到原问题足够精确的解。由定理 2 可知  $\Phi_p(\omega, b)$  是凸函数，因此优化问题 D 的任一局部最优解都是全局最优解。下面给出标准无约束优化算法的子程序。

**基本算法。**

- 1. 给定任一初始点  $(\omega^{(0)}, b^{(0)})$  和正则化参数  $C$ ，令  $p$  为一个充分大的常数；
- 2. 用无约束优化算法子程序进行  $\Phi_p(\omega, b)$  的最小化计算；
- 3. 将最优解  $(\omega^*, b^*)$  代入式(6)，从而得到广义最优线性分类器  $f(\mathbf{x}) = \text{sgn}\{(\omega^{*T} \mathbf{x}) + b^*\}$ 。

**3.2 特征选择问题的数学模型**

给定一个特征子集  $F = \{f_1, f_2, \dots, f_N\}$ ， $N$  是特征集的大小。一个特征子集可以用一个二进制向量表示：

$\mathbf{S} = (s_1, s_2, \dots, s_N)$ ， $s_i \in \{0, 1\}$ ， $i = 1, 2, \dots, N$ ， $s_i = 1$  表示第  $i$  个特征  $f_i$  被选择，反之对第  $i$  个特征  $f_i$  不作选择。把优化的线性支持向量机在给定的特征子集  $\mathbf{S}$  上所具有的性能  $G(\mathbf{S})$  作为目标函数值，则特征选择问题转化为下列优化问题

$$\max_{\mathbf{S}} G(\mathbf{S}).$$

特征选择的求解优化问题  $\max_{\mathbf{S}} G(\mathbf{S})$  可以通过遗传算法和禁忌搜索的混合策略 GATS 来求解。

**3.3 GA 与 TS 的混合搜索策略 GATS**

遗传算法(Genetic Algorithm, GA)是由美国学者 Holland 和他的同事、学生提出的。该算法基于 Darwin 的进化论和 Mendel 的遗传学说，是一种进化搜索算法，已经成为求解任意函数优化问题的强有力的工具<sup>[6]</sup>。其中重组和变异算子是 GA 的两个最重要的组成部分。禁忌搜索(Tabu Search, TS)是另一个著名的启发式搜索算法，最早由 Glover 提出<sup>[7]</sup>，具有记忆功能是 TS 独创的特点之一。开发混合算法的目的是使原算法的优点被保持，弱点被克服或者被削弱，提高算法的力度。Mülenbein 最早把记忆功能引入到 GA<sup>[22]</sup> 中，而 TS 的创始人 Glover 对混合 GA 与 TS 的必要性和可行性进行了理论上的分析和论述<sup>[23]</sup>，被公认为混合 GA 与 TS 的理论基础。在 Glover 理论的基础上，本文提出一种 GA 与 TS 的混合策略 GATS。把 TS 独有的记忆功能引入到 GA 进化搜索过程之中，构造了新的重组算子 TSR。针对 GA 爬山能力差的缺陷，利用 TS 爬山能力强的优点，使用 TS 算法改进 GA 的爬山能力，即把 TS 作为 GA 的变异算子 TSM。GATS 综合了 GA 具有多出发点和 TS 记忆功能和爬山能力强的特点，克服了 GA 爬山能力差的弱点，并保持了 GA 具有多出发点的优势。

**3.3.1 变异算子 TSM 与重组算子 TSR**

TSM(Tabu Search Mutation)与标准变异算子极为相似，首先，TSM 把一个染色体作为输入初始解，经过 TSM 作用，返回一个解作为输出。不同之处在于 TSM 是一个搜索过程，因此需要调用评价函数来确定移动值，并根据移动值和禁忌表  $T$  决定接受哪个移动输出。同样由于 TSM 是一个 TS 搜索过程，在搜索过程中可以接受劣解，因此 TSM 具有强于其它如到位和部分到位算子的爬山能力。设  $x$  是一个染色体，则 TSM 的操作过程如图 2 所示。

```
begin
    t=0; set the best solution x(0)=x; set T;
    while termination condition not satisfied do
        t=t+1;
        move x to x';
        update(x, x(0), tabu list);
    end
```

图 2 TSM 操作过程

TSR(Tabu Search Recombination)算子作为重组算子，使用一个长度为  $T$  的禁忌表，表中记录染

色体的适应值,渴望水平作为父代群体适应值的平均值. 进行 TSR 操作时,首先把子代的适应值同渴望水平相比较,如果渴望水平好,则破禁,即这个染色体进入到下一代中. 如果子代比渴望水平差,但不属于禁忌,也接受这个子代,若属于禁忌,则选择那个最好的父代进入到下一代中. TSR 的重组过程如图 3 所示.

```
begin
  if fitness of  $x$  > average value of population
  then accept  $x$ ;
  else
    if offspring  $x$  is not in tabu list
    accept  $x$ 
  else
    choose the better of two parents to the next generation;
    update tabu list;
end
```

图 3 TSR 重组过程

从 TSR 的重组过程可以看出,具有高适应值的子代进入到下一代的机会是很大的,但是并不是所有的高适应值的子代一定都进入到下一代. 因为 TSR 使用了禁忌表,它可以限制适应值相同的子代出现的次数,因此可使群体中尽可能保持染色体结构的多样性,从而避免算法早熟.

3.3.2 编码方案

对特征子集  $F = \{f_1, f_2, \dots, f_N\}$  中的每一个特征利用二进制进行编码,得到一个码长为  $N$  的二进制串:  $h = h_1 h_2 \dots h_N$ , 这一编码串表示对特征集所做的一次选择,其中  $h_i = 1$  表示第  $i$  特征被选择,所有选择的特征构成一个特征子集.  $H = \{h_1 h_2 \dots h_N \mid h_i \in \{0, 1\}, i = 1, 2, \dots, N\}$  为所有的特征子集的集合,称为个体空间,个体空间的大小为

$$C_N^1 + C_N^2 + \dots + C_N^N = 2^N - 1.$$

3.3.3 适应度定义

大多数基于遗传算法的 wrapper 型特征选择方法中,采用某些分类器模型对所选择的特征集合进行评价,并利用得到的分类精度或分类错误率作为适应度函数. 本文采用优化的线性支持向量机模型作为分类模型. 对于每一个体  $h = h_1 h_2 \dots h_N \in H$ , 将样本数据按随机的原则分为两部分,分别以它们作为训练集和测试集,应用训练集训练支持向量机,然后在测试集上进行验证其分类的正确率(即模型的推广能力),最后计算平均正确分类率. 对于一给定的特征子集,希望做到平均正确分类率尽可能大,故以平均正确分类率作为个体的目标函数值. 设  $\gamma_1, \gamma_2, \dots, \gamma_M$  表示  $M$  类分类问题中的每一类正确分类

率,于是搜索的目的就是寻求极大化问题  $\max_{h \in H} f(h)$  的最优解,其中  $f(h) = \frac{1}{M} \sum_{i=1}^M \gamma_i$ .

遗传算法在执行过程中是按个体的适应度分配选择概率的,适应度好的个体被遗传下一代的可能性大,相反适应度差的个体被遗传到下一代的可能性较小. 通常情况下可选择个体的目标函数值作为其适应值. 考虑到特征选择的目的是选择尽可能小的特征子集,故可对包含较多特征的个体给予一定的惩罚,抑制这些个体选择概率而让包含较少特征的个体有更多的繁殖机会. 本文采用一种改进的适应度函数:  $f'(h) = f(h) - c \times l(h)$ ,  $f'(h)$  是带有惩罚参数  $c$  的适应度函数,  $l(h)$  为个体  $h$  中 1 的位数. 惩罚参数  $c$  的大小根据需要进行选择.

3.3.4 基于无约束优化线性支持向量机和 GATS 的特征选择算法

在 GATS 框架下,得到一种新的基于无约束优化线性支持向量机和 GATS 的特征选择方法 (Feature Selection based on Optimized SVM and GATS, FSOSGT), 算法具体描述如图 4 所示,算法在最优特征子集搜索过程中使用了杰出者记录策略<sup>[24]</sup>.

FSOSGT( $D, F, \delta, N, MaxI$ )  
输入: 数据集  $D$ , 特征集合  $F$ , 阈值  $\delta$ , 群体规模  $N$ , 最大进化代数  $MaxI$   
输出: 最优化特征子集  
1. 初始化: 确定种群规模  $N$  及中止规则(如设置最大进化代数  $MaxI$  或达到的近似解的精度  $\delta$ ); 由特征集合  $F$  随机生成  $N$  个个体子集作为初始种群  $X(0)$ , 记  $X(0)$  具有最高适应值的个体为  $X^*(0)$ (有多个时取其中之一), 置进化代数器  $t = 0$ ;  
2. 个体评价: 采用改进的适应度函数  $f'(h)$  在数据集  $D$  上计算第  $t$  代种群  $X(t)$  中每一个个体的适应度;  
3. 选择: 将选择算子作用于种群  $X(t)$ ;  
4. 繁殖: 将交叉算子, TSR 重组过程和 TSM 操作过程作用于选择后的种群,生成下一代种群  $X(t+1)$ , 记  $X(t+1)$  具有最高适应值的个体为  $A(t+1)$ , 令  
     $X^*(t+1) = \max\{X^*(t), A(t+1)\}$ ;  
5. 终止检验: 如  $X^*(t+1)$  满足终止条件, 则输出  $X^*(t+1)$  作为近似最优解, 终止计算; 否则置  $t = t+1$ , 转步 2.

图 4 FSOSGT 特征选择算法

4 基于特征选择算法的轻量级入侵检测系统

基于特征选择的轻量级入侵检测系统流程见图 5. 流程图的主要部分是特征选择. 首先由搜索策略 GATS 产生一个组随机特征子集  $S_0$  作为初始子集集合,  $S_0$  经过优化的线性支持向量机评估, 得出最优评估结果  $\theta_{best}$ . 申请变量  $S_{best}$ , 它的初始值为

$S_{0\ best}$ ，与之相对应的初始评估值为  $\theta_{best}$ ，这样  $S_{best}$ ， $\theta_{best}$  的初始化过程就结束. 下面进入迭代过程，在每一次迭代过程中，由 GATS 产生的特征子集集合  $S$  均提交给优化的线性支持向量机进行评估，得到最优的评估值为  $\theta$ ，其对应的特征子集为  $S_{best}$ . 如果  $\theta$  大于  $\theta_{best}$ ，则把  $S_{best}$  的值赋给  $S_{best}$ ， $\theta_{best}$  的值等于  $\theta$  的值. 如果出现  $\theta$  值小于  $\theta_{best}$  值的情况，则进入下一次迭代. 当迭代进入到最后阶段，判断预先设定的标准  $\delta$  是否得到了满足，如果迭代达到要求的标准，则迭代结束，把迭代过程中产生的最优特征子集  $S_{best}$  提交给分类器，建立相应的入侵检测模型. 如果预先设定的标准  $\delta$  没有达到，而且迭代的最大次数也没有达到，则进入下一轮迭代. 当特征选择结束之后，在训练集  $TrD$  上用选择的特征子集训练模型  $C$ ，然后在测试集  $TeD$  上测试模型  $C$  的性能. 在本文的实验部分，将对基于  $S_{best}$  建立的入侵检测系统和基于 KDD1999<sup>①</sup> 所有 41 个特征建立的入侵检测系统进行详细的比较和分析，主要从建立模型的时间、检测时间、对已知攻击和未知攻击的检测能力 4 个方面进行对比.

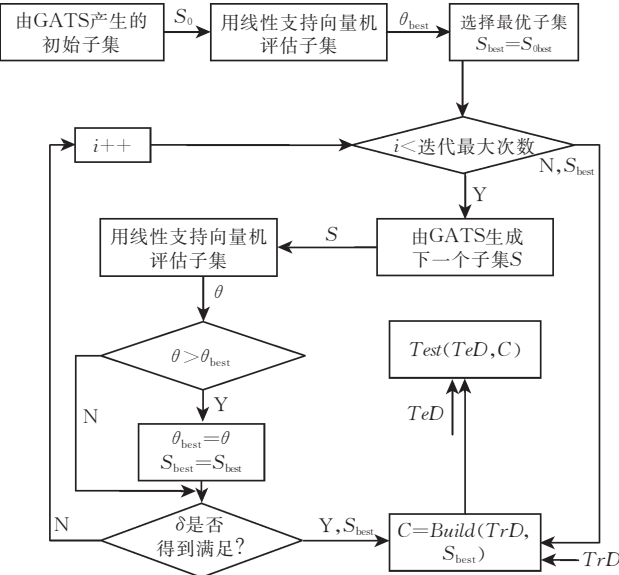


图 5 基于 wrapper 型特征选择算法的轻量级入侵检测系统流程图

5 实验研究

整个实验的流程在图 5 中已经有详细的描述，其中无约束优化线性支持向量机是封装在 GATS 混合策略中. 实验目的有三点：首先是验证本文提出的 wrapper 型特征选择算法能够加快特征选择的速

度；其次验证结合特征选择算法的入侵检测系统与没有特征选择算法的入侵检测系统在系统建模时间、检测时间、检测率上具有更好的性能；最后阐述本文提出的特征选择算法与当前典型的特征选择算法相比在检测率相近的条件下具有更少的建模时间与检测时间. 当前基于特征选择算法的入侵检测系统关注更多的是检测时间<sup>[25]</sup>，而不是检测正确率. 文献<sup>[25]</sup>强调不同的特征选择算法对于入侵检测系统的检测性能是相当的，但是在系统建模时间与检测效率上差异很大，所以在特征选择算法的横向比较中，检测时间显得尤为重要. 本文所有的实验均在 KDD1999<sup>①</sup> 数据集上完成，在实验开始之前，我们对 KDD1999 数据集进行了很多预处理工作，以便满足实验的需要. 实验环境是 Windows 操作系统，因特处理器 1.73GHz，512MB RAM.

5.1 数据集预处理

KDD1999<sup>①</sup> 是关于入侵检测的一个标准数据集，它主要分为训练数据集和测试数据集两部分. 训练集中含有 494019 个实例，每一个实例包含 41 个特征. 我们把训练集按照 DOS 攻击、PROBE 攻击、R21 攻击、U2R 攻击和 NORMAL 5 个部分分类，其中 NORMAL 表示正常数据，不包含攻击. 训练集中 DOS 攻击有 6 种不同的类型，PROBE 攻击 4 种，R21 攻击 8 种，U2R 攻击 4 种. 在整个训练集中 NORMAL 实例的数目是 97278，占训练集总实例数的 19.7%，其它 4 种攻击的实例数占总实例的比例为 DOS，79.2%；PROBE，0.83%；R21，0.23%；U2R，0.001%. 它们的实例数以及每一类含有的具体攻击类型在表 1 中有详细的描述. 从表 1 中可以看出，训练集中含有 22 种不同的攻击类型. 对 KDD1999 测试集，我们也做了同样的分类，并且对每一类攻击分成两个部分：已知攻击和未知攻击. 其中已知攻击表示曾经在训练集中出现过的攻击，未知攻击指在训练集中没有出现过的攻击类型. 测试集包含的攻击类型有 39 种，其中 DOS 有 10 种，已知攻击 6 种，未知攻击 4 种；PROBE 有 6 种，已知攻击 4 种，未知攻击 2 种；R21 有 15 种，已知攻击 8 种，未知攻击 7 种；U2R 有 8 种，已知攻击 4 种，未知攻击 4 种. 详细的攻击类型由表 1 可以看出. 测试集含有 DOS 攻击的实例数为 229853，其中已知攻击的实例数 223298，占 DOS 总实例的 97.1%；未知攻击的实例数 6555，占 DOS 攻击总实例的 2.9%. 其它 3 种攻击已知攻击和未知攻击的实例数在表 1 中有描述.

① KDD Cup 1999 Data. <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>



表 1 KDD1999 训练集与测试集分类统计信息

	详细攻击类型	训练集中的实例数	测试集中的实例数 (攻击百分比)
DOS	smurf, pod, neptune, back, teardrop, land	391458	223298(97.1%)
	apache2, udpstorm, processtable,mailbomb	0	6555(2.9%)
PROBE	ipsweep, nmap, portsweep, satan	4107	2377(57.1%)
	Saint, mscan	0	1789(42.9%)
R2L	ftp_write,guess_passwd, imap, multihop, phf, spy, warezclient, warezmaster	1126	5993(37%)
	named, snmpguess, xlock, xsnoop, sendmail, snmpgetattack, worm	0	10196(63%)
U2R	buffer_overflow, loadmodule, perl, rootkit	52	39(17.1%)
	Xterm, sqlattack, ps, httptunnel	0	189(82.9%)
NORMAL	—	97278	60593(—)
TOTAL	—	494019	311029(—)

5.2 实验方案设计

通过上面小节对 KDD1999 数据集的分析和处理之后,我们知道训练集含有的实例数比较大,为了使实验方便操作,必须对数据集进行取样,使实例数目减少. 我们对训练集中 DOS, PROBE, R21, U2R, NORMAL 分别采取随机取样,保证抽样出来的样本和原样本在分布上保持一致性. 然后把抽样出的 5 个新样本进行组合,形成实验的训练数据集. 形成的 5 个训练集是: NORMAL 和 DOS 混合, NORMAL 和 PROBE 混合, NORMAL 和 R21 混合, NORMAL 同 U2R 混合, NORMAL 同 DOS, PROBE, R21, U2R 混合. 形成的新的 5 个训练集含有的实例数均为 11701. 在取样后的 5 个训练集上应用本文提出的特征选择算法,选出各个训练集对应的特征子集. 然后在每一个训练集上基于所有 41 个特征和选择后的特征子集建立入侵检测模型. 比较基于所有 41 个特征的人侵检测模型和基于选择后的特征的模型在系统建模时间、检测时间、检测已知攻击和未知攻击方面的性能.

为了给出系统在 DOS 攻击、PROBE 攻击、R21 攻击、U2R 攻击以及上述 4 种攻击的混合攻击上检测已知攻击和未知攻击能力的对比,对 KDD1999 测试集从已知攻击和未知攻击两个部分进行随机取样,形成混合攻击、DOS 攻击、PROBE 攻击、R21 攻击和 U2R 攻击的已知攻击和未知攻击测试集. 在特征选择速度上,对采用基于遗传算法和支持向量机的特征选择算法和本文提出的 FSOSGT 算法进行了对比. 最后对 FSOSGT 与其它典型的特征选择算法<sup>[18]</sup>在检测时间,建模时间上进行了比较.

5.3 实验结果与分析

首先在取样后的 5 个训练集上应用本文提出的特征选择算法,特征选择的结果见表 2. 表 2 左栏表

示攻击类型,包括 ALL, DOS, PROBE, R21, U2R 5 种攻击类型,其中 ALL 表示其它 4 种攻击类型的混合. 表 2 的右栏是针对每一种攻击类型选出的特征子集. 如 ALL 攻击选出的特征子集为 3, 5, 23, 32, 数字表示该特征在 KDD1999 的 41 个特征中的排序序号. ALL 攻击指把 DOS, PROBE, R21, U2R 作为一种攻击类型来处理. 冒号右边的单词是冒号左边的数字对应的特征名称,如 3 对应 service.

表 2 针对各种攻击类型选出的相应特征子集

攻击类型	被选出的特征
ALL	3, 5, 23, 32; service, src_bytes, count, dst_host_count
DOS	2, 5, 23, 34; protocol_type, src_bytes, count, dst_host_same_srv_rate
PROBE	1, 3, 5, 6, 23, 35; duration, service, src_bytes, dst_bytes, count, dst_host_diff_srv_rate
R2L	1, 3, 5; duration, service, src_bytes
U2R	1, 3, 5, 14, 32; duration, service, src_bytes, root_shell, dst_host_count

本文提出的特征选择算法 FSOSGT 是 wrapper 型的,把优化的线性支持向量机封装在遗传算法与禁忌搜索相结合的混合策略 GATS 中. Wrapper 型特征选择算法的优点是选择的特征子集质量高,缺点是选择时间长. FSOSGT 应用了新的搜索策略——GATS 和新的评估标准——无约束优化线性支持向量机,提高了特征选择的速度. FSOSGT 与文献[18]基于遗传算法和线性支持向量机的特征选择算法(Feature Selection based on SVM and GA, FSSG)在选择时间上的比较见表 3. 由表 3 可以看出,FSOSGT 具有更快的特征选择速度,例如在 U2R 上,FSSG 的特征选择时间是 1.5h,而 FSOSGT 却只有 1h,是 FSSG 的 67%左右.

表 3 两种不同的特征选择算法的特征选择时间

算法	特征选择时间/h				
	ALL	DOS	PROBE	R2L	U2R
FSSG	1.3	0.5	4	1.5	1.5
FSOSGT	0.8	0.3	3.2	1.1	1.0

由表 2 给出的特征子集,针对每一类攻击类型,分别在 41 个特征和选出的特征子集上建立入侵检测模型.根据不同的阈值调整,对每一类攻击建立多个入侵检测模型,综合评价入侵检测模型在建模时间、检测时间、检测已知攻击和检测未知攻击方面的能力.每一类攻击类型下基于 41 个特征和选择的特征子集的入侵检测模型的平均建模时间和测试时间见表 4.

表 4 五类入侵检测模型在所有特征和选择的特征子集上平均建模时间和检测时间

入侵检测模型	建模时间/s		检测时间/s	
	所有特征	选择特征	所有特征	选择特征
ALL	78	36	18	8
DOS	136	41	22	9
PROBE	245	123	49	29
R2L	317	35	55	8
U2R	193	85	50	18

由表 4 可知,基于特征选择的入侵检测系统无论是建模时间还是检测时间都比未经特征选择的入侵检测系统要小,特别是 R2L 攻击类型,基于所有特征的系统平均建模时间是 317s,而基于特征选择的系统平均建模时间只有 35s,几乎是所有特征建模时间的 11%.从表 4 我们可以知道,基于特征选择的入侵检测系统比起基于所有特征的入侵检测系统效率更高,更有利于建立轻量级的入侵检测系统.

在选择的特征子集上建立的入侵检测模型,不仅在建模时间和检测时间上具有很大的优势,它们在检测已知攻击和未知攻击的能力上也非常突出.下面给出基于所有特征建立的模型和基于选择的特征建立的模型在检测已知攻击和未知攻击能力上的比较.图 6 给出的是在混合攻击训练数据集上,基于 41 个特征建立的入侵检测模型和基于在其上选择的特征建立的模型检测已知攻击和未知攻击的 ROC 曲线图.

由图 6 可知,基于选择特征建立的入侵检测模型在检测已知攻击和未知攻击上具有更高的 ROC 分值,特别是在检测未知攻击方面,优势更加明显.

图 7 是在 DOS 训练集上建立的入侵检测模型,由图可知,基于选择特征的入侵检测模型无论是检测已知攻击还是检测未知攻击的检测率比起基于所有特征建立的入侵检测模型都要高出几个百分点.

特别是检测未知攻击上,平均高出将近 5%.这些表明,在 DOS 攻击上,基于特征选择的轻量级入侵检测系统在检测已知攻击和未知攻击上具有更好的检测率,提高了入侵检测系统的性能.

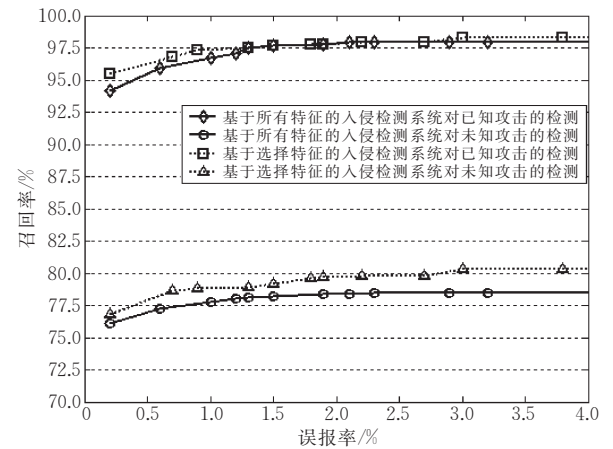


图 6 针对 ALL 数据集的入侵检测模型检测已知攻击和未知攻击的 ROC 曲线图

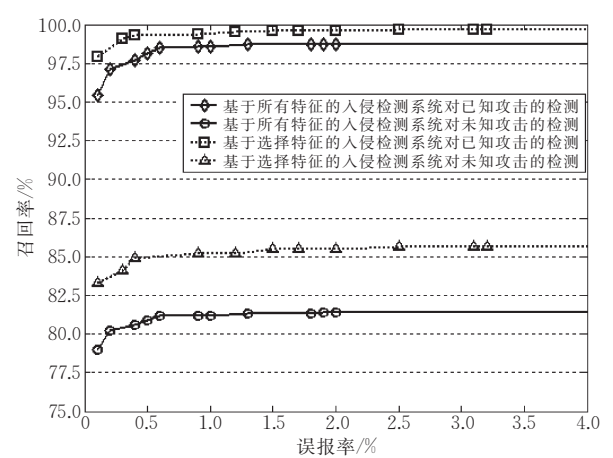


图 7 针对 DOS 数据集的入侵检测模型检测已知攻击和未知攻击的 ROC 曲线图

图 8~图 10 分别给出在 PROBE 攻击训练集、R2L 训练集、U2R 训练集上的基于两种不同的特征集合的入侵检测模型检测已知攻击和未知攻击的 ROC 曲线比较图.从这些比较图可以清楚地看出,基于特征选择算法的轻量级入侵检测模型,不仅能够实现轻量,而且提高了系统检测攻击的能力.特别是对于未知攻击,检测能力的提高更加突出.

从文献[18]可知,FSSG 具有最好的检测率.本文提出的特征选择算法 FSOSGT 与 FSSG 相比,在检测率相近的条件下,比 FSSG 具有更少的建模时间与检测时间.如在建模时间 FSSG 的建模时间是基于所有特征系统的建模时间的 78%,FSOSGT 百分比是 46.7%;在检测时间上 FSSG 的检测时



间是基于所有特征的系统的检测时间的 66.7%，FSOSGT 是 44.5%。从这些数字以及文献[25]可以看出,FSOSGT 比当前典型的特征选择算法 FSSG 具有更小的计算复杂性。

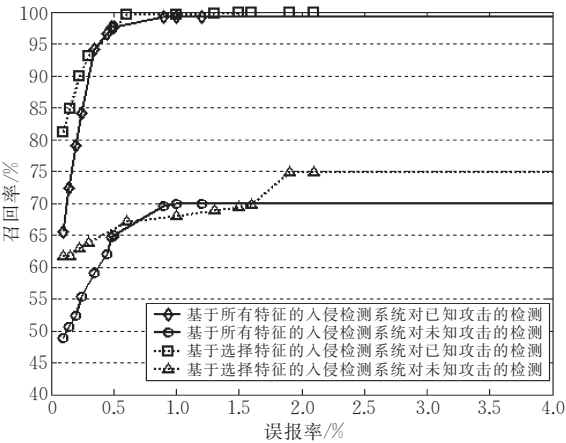


图 8 针对 PROBE 数据集的入侵检测模型检测已知攻击和未知攻击的 ROC 曲线图

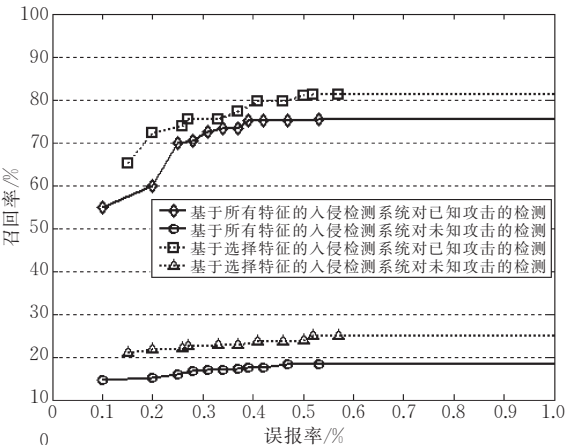


图 9 针对 R21 数据集的入侵检测模型检测已知攻击和未知攻击的 ROC 曲线图

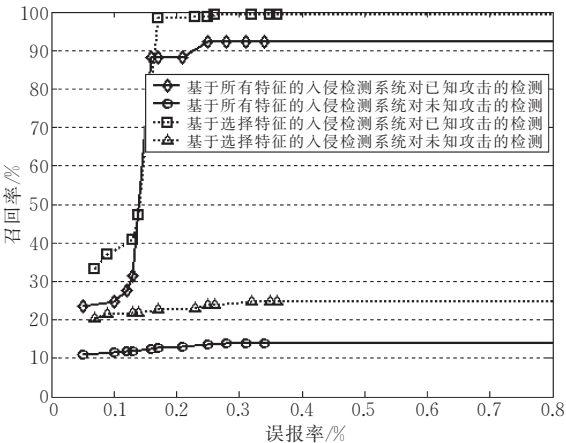


图 10 针对 U2R 数据集的入侵检测模型检测已知攻击和未知攻击的 ROC 曲线图

6 总结和未来的工作方向

现在的学者研究轻量级入侵检测模型主要从两个方面出发:分类器的参数优化和基于数据集的特征选择算法<sup>[26]</sup>. 本文提出了一种 wrapper 型的高效特征选择算法 (Feature Selection based on Optimized SVM and GATS, FSOSGT) 来建立轻量级入侵检测模型. FSOSGT 主要包括:搜索策略——基于遗传算法与禁忌搜索的混合搜索策略 GSTA;评价标准——无约束优化线性支持向量机. 特征选择算法主要分为 filter 型与 wrapper 型两类,wrapper 型与 filter 型相比,选出的特征子集具有更好的性能,但是针对每一个新的特征子集需要更多的计算资源与计算时间去训练分类器,特征选择时间长. FSOSGT 利用无约束优化线性支持向量机提高了训练分类器的速度,同时采用 GA 与 TS 的混合搜索策略,提高了搜索的速度与爬山能力,并且寻找到的特征子集是最优特征子集. 为了验证 FSOSGT 在构建轻量级入侵检测系统中的有效性,本文在 KDD1999 数据集上进行了大量的实验. 实验主要从 5 类攻击入手,所有的 ALL 攻击、DOS 攻击、PROBE 攻击、R21 攻击、U2R 攻击. 针对每一类攻击,建立基于全部 41 个特征和选出的特征子集的两类入侵检测模型. 对这两类入侵检测模型从平均建模时间、平均检测时间、检测已知攻击能力和检测未知攻击能力四个方面进行了比较. 实验结果表明,FSOSGT 具有更小的建模时间与检测时间,能够很好的满足入侵检测系统对实时性的要求;同时,它在检测已知攻击和未知攻击能力上具有更高的检测率,特别是在未知攻击的检测上. 我们未来工作方向主要集中在通过提高搜索策略和评价标准来提高特征选择算法的效率,并且在实际的工程产品中应用我们的特征选择算法 FSOSGT.

参 考 文 献

[1] Forres S, Perelson A S, Allen L et al. Self-nonsel self discrimination in a computer//Proceedings of the 1994 IEEE Symposium on Research in Security and Privacy. Los Alamitos: IEEE Computer Society Press, 1994: 120-128

[2] Kohavi R, John G H. Wrappers for feature subset selection. Artificial Intelligence, 1997, 97(1-2): 273-324

[3] Park Jong Sou, Shazzad K M, Kim D S. Toward modeling lightweight intrusion detection system through correlation-based hybrid feature selection//Feng D, Lin D, Yung M eds. Proceedings of the CISC. Heidelberg: Springer-Verlag, 2005: 279-289

- [4] Jain A K, Zongker D. Feature selection; Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1997, 19(2): 153-158
- [5] Kudo M, Sklansky J. Comparison of algorithms that select features for pattern classifiers. *Pattern Recognition*, 2000, 33(1): 25-41
- [6] Holland J. *Adaptation in Natural and Artificial Systems*. Ann Arbor: The University of Michigan Press, 1975
- [7] Glover F. Future paths for integer programming and links to artificial intelligence. *Computers and Operations Research*, 1986, 13(4): 533-549
- [8] Vapnik V. *The Nature of Statistical Learning Theory*. New York: Springer Verlag, 1995
- [9] Grandvalet Y, Canu S. Adaptive scaling for feature selection in SVMs//*Advances in Neural Information Processing Systems*. Cambridge, Massachusetts: MIT Press, 2003, 15: 553-560
- [10] Cao L J, Chua K S, Chong W K, Lee H P, Gu Q M. A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine. *Neuron Computing*, 2003, 55(1-2): 321-336
- [11] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Machine Learning*, 2002, 46(1-3): 389-422
- [12] Li Xing-Si. An efficient method for nonlinear minmax problems. *Chinese Science Bulletin*, 1991, 36(19): 1448-1451(in Chinese)  
(李兴斯. 非线性极大极小问题的一个有效解法. *科学通报*, 1991, 36(19): 1448-1451)
- [13] Li Xing-Si. An efficient method for a sort of non-differential optimization problems. *Science in China (Series A)*, 1994, 24(4): 371-377(in Chinese)  
(李兴斯. 一类不可微优化问题的有效解法. *中国科学(A辑)*, 1994, 24(4): 371-377)
- [14] Lee Y, Mangasarian O. RSVM: Reduced Support Vector Machines First SIAM International Conference on Data Mining. Wisconsin: University of Wisconsin, 2000: 350-366
- [15] Kim D, Nguyen H-N, Ohn S-Y, Park J. Fusions of GA and SVM for anomaly detection in intrusion detection system//*Advanced in Neural Networks*. Lecture Notes in Computer Science 3498. Springer-Verlag, 2005: 415-420
- [16] Park J S, Sazzad K M, Kim D S. Toward modeling lightweight intrusion detection system through correlation-based hybrid feature selection//Feng D G ed. *Proceedings of the Information Security and Cryptology*. Lecture Notes in Computer Science 3822. Springer-Verlag, 2005: 279-289
- [17] Ribeiro A H, B M. Model selection for kernel based intrusion detection systems//*Proceedings of the International Conference on Adaptive and Natural Computing Algorithms*. Coimbra, Portugal, 2005: 458-461
- [18] Chen You, Cheng Xue-Qi, Li Yang, Dai Lei. Lightweight intrusion detection systems based on feature selection. *Journal of Software*, 2007, 18(7): 1639-1651(in Chinese)  
(陈友, 程学旗, 李洋, 戴磊. 基于特征选择的轻量级入侵检测系统. *软件学报*, 2007, 18(7): 1639-1651)
- [19] Tang Huan-Wen, Zhang Li-Wei, Wang Xue-Hua. A maximum entropy method for a sort of constrained non-differentiable optimization problems. *Mathematica Numerica Sinica*, 1993, 15(3): 268-275(in Chinese)  
(唐焕文, 张立卫, 王雪华. 一类约束不可微优化问题的极大熵方法. *计算数学*, 1993, 15(3): 268-275)
- [20] Tang Huan-Wen, Zhang Li-Wei. A maximum entropy algorithm for convex programming. *Chinese Science Bulletin*, 1994, 39(8): 682-684(in Chinese)  
(唐焕文, 张立卫. 凸规划的极大熵方法. *科学通报*, 1994, 39(8): 682-684)
- [21] Zhang Zhi-Hua, Zheng Nan-Ning, Shi Gang. A maximum entropy clustering method and its global convergence analysis. *Science in China (Series E)*, 2001, 31(1): 59-70(in Chinese)  
(张志华, 郑南宁, 史罡. 极大熵聚类算法及其全局收敛性分析. *中国科学(E辑)*, 2001, 31(1): 59-70)
- [22] Müblenbein H. *Parallel Generic Algorithms in Combinatorial Optimization*. Computer Science and Operations Research (Edited by Osman Balci). Oxford: Pergamon Press, 1995
- [23] Glover F, Kelly J, Laguna M. Genetic algorithms and tabu search: Hybrids for optimizations. *Computers and Operations Research*, 1995, 22(1): 111-134
- [24] Xu Zong-Ben, Zhang Jiang-She, Zheng Ya-Lin. *Bionics in Intelligent Computation: Theory and Algorithm*. Beijing: Science Press, 2003(in Chinese)  
(徐宗本, 张讲社, 郑亚林. *计算智能中的仿生学: 理论与算法*. 北京: 科学出版, 2003)
- [25] Nigel Williams, Sebastian Zander, Grenville Armitrage. A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification. *ACM SIGCOMM Computer Communication Review*, 2006, 36(5): 5-16
- [26] Chen You, Li Yang. Survey and taxonomy of feature selection algorithms in intrusion detection system//Lipmaa H et al. eds. *Proceedings of the Conference on Information Security and Cryptology*. Lecture Notes in Computer Science 4318. Heidelberg: Springer-Verlag, 2006: 153-167



**CHEN You**, born in 1981, Ph.D. candidate. His major research interests include network and information security, data mining.

major research interests include complex networks, information retrieval and information security.

**LI Yang**, born in 1978, Ph.D. candidate. His research interests include network security, intrusion detection techniques based on data mining and machine learning methods, etc.

**CHENG Xue-Qi**, born in 1971, Ph.D., professor. His major research interests include high performance software, information security, and intelligent information process.

**SHEN Hua-Wei**, born in 1982, Ph.D. candidate. His

## Background

Intrusion detection system (IDS) deals with huge amount of data which contains irrelevant and redundant features causing slow training and testing process, higher resource consumption as well as poor detection rate. Existing studies to build lightweight IDS have proposed two main approaches: parameters optimization of classification algorithms and feature selection of audit data. Feature selection is one of the key topics in IDS. Feature selection involves finding a subset of features to improve prediction accuracy or decrease the size of the structure without significantly decreasing prediction accuracy of the classifier built using only the selected features. Methods for feature selection have been essentially divided into two categories: filter methods and wrapper methods. Wrapper methods generally perform better than filter methods, but they involve some more computational complexity and require more execution time than the filter methods. Some researchers have proposed hybrid feature selection methods which combine wrapper and filter methods. Howev-

er, the number of selected features is large and the performances of intrusion detection system which based on hybrid feature selection are not perfect. Current research results show that wrapper feature selection algorithm performs better than other two methods, except its computational complexity. Therefore, this paper proposes a novel wrapper-based feature selection algorithm to build lightweight IDS. The approach is able not only to solve the computational complexity but also to guarantee high detection rates. The researches of this paper are supported in part by the National Basic Research Program (973 Program) of China under grant No. 2004CB318109 and the National Information Security Project of China under grant No. 2005C39. The former project is to develop a security analysis of network systems. The latter is to build lightweight IDS. The work in this paper is part of the research on building lightweight IDS, trying to use novel feature selection algorithm toward building lightweight IDS.