基于构造性核覆盖算法的异常入侵检测

周鸣争, 楚 宁, 强 俊
(安徽工程科技学院计算机科学与工程系, 安徽芜湖 241000)

摘 要: 将构造性核覆盖算法引入入侵检测研究中, 提出了一种基于构造性核覆盖的异常入侵检测算法, 用于监控进程的非正常行为. 首先分析了核覆盖分类算法应用于入侵检测的可能性, 然后具体描述了核覆盖算法在异构数据集下的推广, 提出了基于核覆盖的异常入侵检测模型. 并以 sendmail 系统调用序列数据集为例, 详细讨论了该模型的工作过程. 最后将实验仿真结果与其它方法进行了比较, 结果表明, 该方法的检测效果优于同类的其它方法.

关键词: 异常入侵检测; 核覆盖; 异构数据集; 系统调用
中图分类号: TP309 文献标识码: A 文章编号: 0372-2112 (2007) 05-0862-06

An Anomaly Intrusion Detection Based on Constructive Kernel Covering Algorithm

ZHOU Ming-zheng, CHU Ning, QIANG Jun
(Department of Computer Science and Engineering, Anhui University of Technology and Science, Wuhu, Anhui 241000, China)

Abstract: Constructive Kernel Covering Algorithm is inducted into intrusion detection and an anomaly intrusion detection. This method based on Constructive Kernel Covering Algorithm is proposed which applied to monitoring the abnormal behavior of processes. Firstly, this paper analyzes the possibility of Kernel Covering Algorithm applied to intrusion detection; Secondly, the Kernel Covering Algorithm generalized on heterogeneous datasets is described, and a model of anomaly intrusion detection based on Kernel Covering Algorithm is proposed. Then we illustrate the sendmail system call sequence dataset and discuss the working process of this model in detail; Finally, the emulation result is compared with other methods. The result indicates that this method is superior to other similar methods.

Key words: anomaly intrusion detection; constructive kernel covering algorithm; heterogeneous datasets; system call

1 引言

随着计算机和网络技术应用的日益普及, 计算机网络安全越来越受到人们的重视. 入侵检测(intrusion detection)作为网络安全研究的重要内容, 它是通过监测和分析网络流量、系统审计记录等, 发现和识别系统中的入侵行为和入侵企图, 给出入侵报警, 以便系统管理员采取有效的措施, 弥补系统漏洞和填补系统功能, 已引起了国内外学者的广泛关注.

入侵检测方法一般可分为两大类: 滥用入侵检测(misuse detection)和异常入侵检测(anomaly detection). 滥用入侵检测的基础是建立黑客攻击行为的特征库, 采用特征匹配的方法确定攻击事件, 其优点是检测误报率低, 速度快, 但通常不能发现攻击特征库中没有的实现指定的攻击行为, 所以也无法检测出新的攻击. 异常检测是通过建立用户正常行为模型, 以是否显著偏离正常模型为依据进行入侵检测, 它有一定的误报率. 但它可根据当前的系统行为用正常模型的相似度判断是否为攻击, 有可能发现新的攻击行为. 这两种方法都需要对用户行为进行建模, 滥用检测需要建立攻击行为模式, 异常检测需要建立正常用户的行为模式. 基于机器学习的用户行为建模始终是入侵检测系统(Intrusion Detection System, IDS)的一个重要研究课题, 以往的研究者在 IDS 研究中引入了各种机器的学习方法, 如神经网络[1]、遗传算法[2,3]、HMM[4]、模糊综合评判[5]等, 但这些方法多是基于样本数目趋于无穷大假设的, 并且对数据的规律性要求较高, 在 IDS 中能够获得的数据往往呈现出多变性、小样本和高维性, 较难满足上述这些算法的要求, 使得检测结果实用性较差, 误报率较高.

近几年来成为机器学习研究热点的核函数方法, 是一种比较完备的从小样本数据中寻找规律的系统方法, 主要用于解决有限样本的模式识别分类问题. 将 SVM 核函数方法应用于入侵检测系统也有相关的研究[6], 但对 SVM 核函数方法中存在的核函数参数难以确定、计

收稿日期: 2006-04-29; 修回日期: 2006-09-15
基金项目: 国家自然科学基金(No. 60574028); 安徽省教育厅自然科学重点项目(No. 2004kj033zd)

算复杂性高的问题均未获得突破性的解决. 本文将构造性核覆盖算法(Constructive Kernel Covering Algorithm, CKCA)[7]应用于入侵检测中, 针对 IDS 中经常出现的异构数据集情况, 构造了基于异构距离定义的 RBF 形核函数, 将核函数方法与构造性学习的覆盖算法相融合, 提出了一种异构数据集下的核覆盖的网络入侵检测方法, 克服了 SVM 核函数方法的缺点, 保证了在先验知识不足和小样本的情况下, 系统仍有较好的分类正确率和较小的运算量, 从而使得整个 IDS 具有较好的检测性能.

2　构造性核覆盖算法与入侵检测

2.1 构造性核覆盖算法

2.1.1 覆盖算法与 SVM 的核学习方法

现有的前馈神经网络在求解模式识别问题时, 其中心思想是建立样本和其类别的映射, 然而其训练方法是基于预先给定的评价函数的极小化, 其本质也是用形式和数目预先确定的多个函数(即隐层单元的输出函数)的组合去逼近这一映射. 但它没有直接从样本数据本身入手来推测有关性质.

1998 年, 张铃等人提出了一种新的 *M-P* 神经元的几何意义解释[8] ——"球"领域模型, 从而将神经网络的训练问题转化为求解"球领域"模型的最小几何覆盖问题, 并以此给出了一种设计神经网络(作为分类器)的覆盖算法.

M-P 神经元是一个 n 个输入、单输出的元件, 其输入与输出的关系为:

$$Y=\mathrm{Sgn}(\sum w_i x_i-\alpha) \tag{1}$$

若令 $\sum w_i x_i-\alpha=0$, 则此式表示为一个超平面方程. 于是从几何意义上可将神经元看成是一个空间分类器, 即落在正半空间的点对应的输出为 1, 落在负半空间上的点对应的输出为-1. 若设输入样本的长度相等, 即输入样本分布在 $n+1$ 维空间的某个球面 S^n 上(其中心点在原点, 半径为 R). 那么这时($W^* x-\theta>0$(其中 W 是权向量, θ 是阈值), 就表示球面上落在由超平面 P(其方程为($W^* x-\theta=0$)所分割的正半空间的部分, 这个部分恰好是球面上的某个"球形领域". 若取 W 与 x 等长, 则这个"球形领域"的中心恰好是 W, 其半径为:

$$r(\theta)=\cos^{-1}(\theta/R^2) \tag{2}$$

若令 $\sigma'(x)=\begin{cases}1, & \text{当 } x>0 \text{ 且取神经元的激励函数为 } \sigma'(W^* x-\theta),\\ 0, & \text{其他}\end{cases}$

则一个神经元的激励函数正好是它所代表的球面上"球形领域"的特征函数, 这样, 就将神经元与球面上的球形领域对应起来, 利用神经元的这种几何意义, 就能非常直观地进行神经网络的各种研究.

由上面给出的神经元的几何意义得知, 构造一个网络, 对给定的样本集能进行符合要求的分类, 等价于求出一组领域, 对给定样本集 K 中的点, 能按分类的要求用领域覆盖将它们分割开来. 这样, 就将神经网络的最优设计问题转化成某种求最优覆盖的问题. 同时也将原先基于搜索机制的学习方法转变成构造性的学习方法. 从而为处理海量数据提供了一种切实可行的方法.

当给定的输入向量的长度不相等时, 可用下面给出的方法, 将它变换成长度相等的情况. 设输入的定义域为 n 维空间中的有界集合 D, 令 S^n 是 $n+1$ 维空间中的 n 维的超球面, 作变换 $T: D \to S^n$, $x \in D$,

$$T(x)=(x, \sqrt{(d^2-|x|^2)}) \tag{3}$$

其中 $d \geqslant \max\{|x|, x \in D\}$.

覆盖算法通过反复迭代求取最优的"球领域"覆盖中心, 以使得球领域覆盖数最少. 该算法的主要特点是, 在对具体数据处理过程中, 把求解样本集 S 的 k 类分类问题转化成在样本空间构造覆盖簇 $\{C_i\}$, 使每个覆盖 C_i 只盖住同一类点. 设已求得覆盖簇 C_1, C_2, C_3, …, C_n, 取三层神经网络, 隐层取 n 个神经元, 每个神经元为一个覆盖, 第 i 个神经元的激励函数为 C_i 的特征函数. 输出层取 k 个神经元, 第 i 个神经元的输入为覆盖第 i 类点的覆盖的输出, 其激励函数为或门. 这样的三层网络, 就可对 S 进行分类, 解决了神经网络结构难以确定的问题, 具有一次处理多分类问题的优点. 但该算法存在有以下两点不足:

(1)最终覆盖的质量受先验知识和初始种子样本的选择以及样本分布的影响较大;

(2)该算法虽然在某种意义上可以使得神经网络规模最小, 该算法相对复杂, 实现比较困难.

Vapnik 提出的支持向量机(SVM)学习算法[9], 对线性可分情况给出一个用规划方法解得的最大间隔解. 对线性不可分情况提出用核函数, 将原问题映射到高维空间, 然后在这个新空间中求取最优线性分类面, 其所求得的分类函数形式上类似于一个神经网络, 其输出是若干中间层节点的线性组合, 而每一个中间层节点对应于输入样本与一个支持向量的内积. 它不像传统方法, 先试图将原输入空间降维(即进行特征选择和特征变换), 而是设法将输入空间升维, 以求在高维空间中, 使问题变得线性可分.

从几何意义上看, SVM 中求最优分类面问题就是在对应的核函数类中, 求得一个其零值等高线是两类的分界线以及样本集到边界的距离最大的函数. 也就是求划分边界线的问题. 如果能将边界线"附近"的点分开, 那么其它的点就自然而然地被分开. 如果沿零等

高线的两边, 以最大间隔为宽度, 划一条与之"平行"的线, 即得到一条"边界河", 那么支持向量就必落在边界河的河沿上. 由于不同的核函数所构成的"边界河"是不同的, 因此不同类型的 SVM 利用的支持向量也是不同的.

支持向量机方法是到目前为止, 统计学习理论最成功的实现, 它是在统计学习理论的 VC 维理论和结构风险最小原理基础上, 尽量提高学习机的泛化能力, 能够较好地解决小样本、非线性和高维数的非线性分类问题. 但是对 SVM 方法中存在的核函数选取问题(特别是核函数的参数选取问题)以及因要求二次规划而引起计算量大的问题, 目前均未获得突破性的解决.

通过上述的分析, 可以看出覆盖算法与核函数方法有如下关系:

(1)对于二分类问题, 设覆盖法得到的输出为 $F(y)$, 令 $K(x, y)=\langle T(x), T(y)\rangle$, 则 $K(x, y)$ 是与覆盖算法对应的核函数. 这个映射有如下特点: 一个覆盖邻域中的点被映射到空间 Z 中的同一方向, 而不同的覆盖邻域则被映射成相互正交的方向. 这个性质使求解问题大为简化, 而核函数法得到的支持向量集, 与每个支持向量相对应的方向则并不一定正交.

(2)核函数方法也可看成是一种特殊的覆盖算法, 只要对每个支持向量 x_i 取覆盖的功能函数为 $\sigma(K(x, x_i))$, 其中:

$$\sigma(x)=\begin{cases}1, & \text{当 } x>0\\ 0, & \text{其他}\end{cases}$$

(3)利用核函数方法求得的是最大间隔解, 而覆盖算法在求解过程中只有局部求优的过程.

2.1.2 构造性核覆盖算法

基于上述核函数与覆盖算法的等价性特点, 在覆盖算法中引入核函数[7]. 首先, 任取一核函数 $K(x, y)=\langle T(x), T(y)\rangle$ 做以下变换 $T: D\rightarrow Z$, $x\in D$; 其中 D 为输入的定义域为 n 维空间的有界集合, 共有 p 个样本. 这种变换就是将 D 上的点映射到 P 维核空间上, 记核空间的输入集为: P_t, $t=1, 2, \cdots, p$. 在核空间中, 不妨设输出集 Y 的前 k 个值均不相同. 令所有输出为 $y_j(j\leqslant k)$ 的样本标号的集合为 I_j(即 $I_j=\{I\mid y_i=y_j\}$), 其对应的输入集合记为 P_j, $j=0, 1, 2, \cdots, k-1$. 经过上面的一系列初始化后, 即可开始求取一批核空间中的覆盖 $\{C_j^{(i)}, j=1, 2, \cdots, s-1; i=1, 2, \cdots, p\}$. 令 $C_j=\bigcup C_j^{(i)}$, $i=1, 2, \cdots, p$, 则每个 C_j 表示一个类别的所有覆盖. 因此构造性核覆盖算法的步骤为:

Step 1 在样本集中任取一个尚未被覆盖的点 x_j, 使得 $x_j\in P_t$, 按式:

$$d_j^{(1)}=\min_{m\in I_j}\{K(x_j, x_m)\} \tag{4}$$

$$d_j^{(2)}=\max_{m\in I_j}\{K(x_j, x_m)\mid K(x_j, x_m)<d_j^{(1)}\} \tag{5}$$

$$d_j=[d_j^{(1)}+d_j^{(2)}]/2 \tag{6}$$

$$\theta_j=[d_j^{(1)}-d_j^{(2)}]/2 \tag{7}$$

计算, 根据 x_j 和 d_j 构造一个覆盖 $C_j^{(i)}$, 该覆盖的中心为 x_j, 覆盖半径 $R=d_j$, 分类间隔为 d_j.

Step 2 $C_j^{(i)}$ 求出后, 将 P_t 中所有的已被 $C_j^{(i)}$ 覆盖的点从 P_t 中删除, 再在 P_t 中选择一个 $x_j(j\in I_j)$, 重复第一步操作, 直到所有的 $x_j\in I_j$ 均已被删除为止. 这样, 便构造出一个类的所有覆盖领域.

Step 3 对所求出的中心、半径分别为 $x_1, x_2, \cdots, x_m$ 和 $d_1, d_2, \cdots, d_m$ 的覆盖领域. 令

$$K(x, y)=\exp(-\beta|x-y|^2/d_i^2) \tag{8}$$

其中, d_i 表示以 x 为中心的领域的半径, 求二次规划问题:

$$\max w(\alpha)=\sum_{i=1}^{m}a_i-\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j(K(d_i, d_j)+K(d_j, d_i)/2) \tag{9}$$

$$\text{s.t.}\quad \sum_{i=1}^{m}\alpha_i y_i=0,\ \alpha_i\geqslant 0,\ i=1, 2, \cdots, m \tag{10}$$

得到最优解:

$$\alpha^*=\{\alpha_1, \alpha_2, \cdots, \alpha_m\}$$

Step 4 用 α^* 构造超平面:

$$f(x)=\sum_{i=1}^{m}\alpha_i y_i K(d_i, x) \tag{11}$$

其判别函数为:

$$F(x)=\text{Sign}(f(x)+b_0) \tag{12}$$

其中 b_0 为决策阈值.

Step 5 对样本进行分类. 对每一个样本, 计算 $f(x)$ 的值, 若 $f(x)>0$, 则 x 属于正类, 若 $f(x)<0$, 则 x 属于负类, 若 $f(x)=0$, 称 x 被拒识. 在数据有噪声或对精度要求很严格要求时, 可以设定一个阈值 ε, 当 $|f(x)|<\varepsilon$ 时认为 x 被拒识, 这样可以减少误差所造成的损失.

与 SVM 算法相比, 该算法具有如下的一些特点:

(1)对任意给定的样本集, 该算法能构造出一次就可准确划分样本集的核函数.

(2)在对 $C_j=\bigcup C_j^{(i)}$ 求优时, 求和项只对覆盖集取和, 而不是对所有样本点取和. SVM 中最后其求和式只对支持向量集取和, 但那是在求优之后才得出的, 其在求解时是对所有样本点求和. 一般覆盖个数要比样本点的个数少得多, 这就是 SVM 计算量过大得原因. 故该算法计算量比 SVM 少.

(3)CKCA 通过覆盖后, 在已求得一个允许解的情况下, 可再在这个基础上求解最优, 这比从任给的初始点开始求最优解要快得多.

2.2 构造核覆盖算法应用于入侵检测的可行性

从方法上讲, 滥用入侵检测的核心是攻击行为模式的正确表达和快速识别. 异常入侵检测的核心是用户正常行为模式的建立以及正常模式和异常模式的识别和分类, 这些都是典型的模式识别问题. 因此广义上讲入侵检测是属于模式识别的范畴. 从本质上讲, 入侵检测可以看作是一个分类问题, 也就是通过检测把正常数据和异常数据分开, 但是 IDS 中需要分类的数据更加复杂, 常常体现为高维、小样本和不可分性. Vapnik 提出的支持向量机(SVM)学习方法, 对线性可分情况给出一个用规划方法解得的最大间隔解; 对线性不可分情况提出利用核函数, 将原问题映射到高维空间, 然后按线性可分的情况进行处理. 实践证明在一般情况下, 核函数法是有效的, 但核函数中的参数难以确定, 计算复杂性高; 构造性覆盖学习算法根据训练样本构造性地设计分类网络, 运行效率高, 但存在所的的分界面零乱、测试计算量大的缺点. 核覆盖算法将核函数法与构造性学习的覆盖相融合, 克服了核函数法和构造性覆盖算法的缺点, 具有运算速度快、精度高、鲁棒性强的优点, 可广泛用于小样本数据的分类, 而且对数据维数不敏感, 即由有限的训练集样本得到小的误差仍然能够保证对独立的测试集保持小的误差. 同时也可用于密度估计和孤立点的发现, 既不均衡数据集中无监督的异常检测问题. 因此核覆盖算法适合于入侵检测中对高维异构不均衡数据集进行分类和异常发现, 将其应用于入侵检测是可行的.

2.3 基于构造性核覆盖的入侵检测系统

2.3.1 异构数据集上的核覆盖算法

构造性核覆盖算法中核函数是在输入空间为内积空间下推导出来的, 而异构数据集上通常无法定义内积. 为此我们在文献[10, 11]的基础上, 引入了基于异构数据集上的异构距离的核函数, 将构造性核覆盖算法在异构数据集上进行了推广.

设 $x, y \in X$, 则 x, y 之间的异构距离可用以下方式定义

$$H(x, y) = \sqrt{\sum_{a=1}^{m} d_a^2(x_a, y_a)} \tag{13}$$

其中:

$$d_a(x, y) = \begin{cases} 1, \text{如果 } x_a \text{ 或 } y_a \text{ 未知;} \\ normalized\text{-}vdm_a(x, y), \text{如果第 } a \text{ 个属性是离散值;} \\ normalized\text{-}diff\,a_a(x, y), \text{如果第 } a \text{ 个属性是连续值} \end{cases} \tag{14}$$

$$normalized\text{-}diff\,a_a(x, y) = \frac{|x - y|}{4\sigma_l} \tag{15}$$

$$normalized\text{-}vdm_a(x, y) = \sqrt{\sum_{c=1}^{c} \left| \frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}} \right|^2} \tag{16}$$

σ 为数据集上第 l 个属性的方差, $N_{a,x}$ 为数据集 X 上所有数据第 a 个属性取值为 x_a 的数据个数, $N_{a,x,c}$ 为数据集 X 上的所有数据第 a 个属性取值为 x_a, 且输出类别为 C 的数据的个数, C 是数据集的所有输出类别.

这种距离定义对不同的属性采用不同的距离定义方式, 最终的距离用各个属性分量的欧式距离来获得, 充分表达了数据集之间的相似性, 更好度量数据之间的距离而且计算比较简单、高效.

通常入侵检测中使用的数据集为一异构数据形, 这就需要构造一个特殊的核函数将这种向量数据集变换到一个内积空间上, 然后实现各种算法. 我们对 RBF 核函数进行了改进来完成这种变换, 用输入异构数据集上的异构距离来代替范数, 再通过 RBF 核函数定义将输入数据集映射到一个特征空间, 构成一个满足构造性核覆盖的条件, 用这种改进的 RBF 核函数做构造性核覆盖的分类. 最终的核函数定义为:

$$K(x, x_i) = \exp(-H(x, x_i)/\delta^2) \tag{17}$$

其中, H 为异构距离函数.

2.3.2 基于构造性核覆盖的入侵检测系统

基于构造性核覆盖算法的入侵检测系统主要由审计数据预处理器, 构造性核覆盖分类器两部分组成, 如图 1 所示:

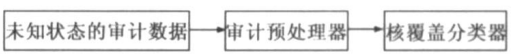


图 1　基于构造性核覆盖算法的入侵检测系统

审计数据预处理系统用于对大量的系统审计记录进行处理和变换, 由于其构造性核覆盖的分类器只能对维数相同的数字向量进行分类, 但有时系统审计数据中的数据不但长度不尽相同, 而且很有可能不是数字类型, 所以必须将审计数据按距离函数的定义转换成构造性核覆盖算法分类器能够处理的数字向量; 核覆盖分类器对这些数字向量进行分类, 产生分类结果, 这些结果可以直接作为整个入侵检测系统的输出.

系统工作分为训练阶段和检测阶段, 在训练阶段, 根据已知的正常审计数据和异常审计数据按上述构造性核覆盖算法计算出相关的参数, 构建出相应的核覆盖分类器; 在检测阶段, 预处理器先将未知状态的审计数据处理成数字向量的形式, 然后通过核覆盖分类器, 根据式(12)对输入数字向量进行分类, 并将分类结果提交给决策系统作出最后的判断.

3　实验测试及分析

为便于比较, 本文测试采用美国新墨西哥大学计算机科学系 Stephanie Forrest 教授领导的实验室收集的

系统调用序列数据集 Sendmail[12]. 系统调用序列是入侵检测中使用比较普遍的一种审计数据, 一条系统调用序列就是一个进程在运行过程中发出的所有系统调用的顺序序列, 在这个序列可以通过系统中的应用程序(比如 strace) 获得. 一个正常行为可以由其执行迹局部模式, 即系统调用短序列来描叙其程序执行代码具有相对的稳定性. 在异常行为中, 可能出现和正常情况有一定差别的系统调用短序列. 所以, 某个进程发出的执行迹是正常的还是异常的就转化为识别执行迹中的短序列是正常的还是异常的, 进而确定该进程是否有安全方面的问题.

3.1 实验数据的预处理

Sendmail 进程常用在 Unix 系统下, 完成接收和发送邮件工作, 是一个重要的进程, 每个系列数据文件有两列整数组成, 第 1 列表明进程 ID 号, 另一列则是代表某个系统调用号. 这些代号可以通过一个索引文件转化成具体的系统调用名称, 例如: 代号“5” 表示名称为“Open” 的系统调用进程标识符相同的系统调用构成一个进程的执行迹. Sendmail 进程可以分叉, 它的子进程产生的系统调用序列被单独跟踪, 但它们也被包括在当前 Sendmail 的序列中, 用不同的 Pid 加以区分, 由于其审计数据已是数字序列, 预处理的主要目的是得到该执行迹的系统调用短序列. 实验时先对实验数据进行了归一化处理: 首先, 对具有不同进程号的调用序列进行分离. 再对原始序列切割成某一给定长度的序列段, 根据 Lee[13] 从信息论的角度研究的选择结果, 认为最适合的系统调用短序列长度为 6~ 7. 本实验中, 数据预处理采用长度为 6 的窗口在程序执行迹上滑动来得到执行迹的短序列, 通过对正常的系统调用执行迹扫描, 可以得到正常的系统调用短序列样本. 对异常的执行迹扫描时, 会得到一组既有正常短序列又有异常短序列的系统调用短序列列表, 通过比较, 就构成了异常短序列样本. 最后对归一化的数据随机抽取约 10% 作为训练集, 剩余的 90% 作为测试集.

3.2 实验结果及讨论

通过训练, 得到核覆盖算法的相关参数后, 首先用长度为 6 的滑动窗口对测试集的执行迹进行扫描, 产生一组系统调用短序列. 将这些系统调用短序列作为基于构造性核覆盖分类器的输入, 利用式(12) 可以得到相应执行迹的状态. 表 1 为本文训练数据和检测数据的分配情况. 表 2 为其本文方法的仿真结果以及与 Forrest 等在文献[14] 中和 Lee 等在文献[15] 中得出研究结果的序列异常度比较.

通过比较, 可以看到, 本文的方法具有一定的优势. 从正常异常序列的差值(异常序列中的最小异常度减去正常序列的异常度) 来看, Forrest 方法给出的结果是 5.66, Lee 的方法给出的结果是 24.27, 而我们的方法得到的结果为 23.66. 我们的方法略好于 Lee 的方法. 能够准确地将进程的正常运行状态同异常运行状态区分开来. 但若同时考虑正常序列异常度的实际值, 我们的方法是 2.24 而 Lee 方法是 4.41, Forrest 将所有的正常序列用于训练, 正常序列的异常度必然为 0, 不能说明其准确度高. 基于上述的讨论, 可以看出在 3 种方法中, 本文方法对进程正常行为的表达更准确, 能最有效地将正常和异常执行序列区分开来, 在给定的误报水平上, 可以降低阈值, 减少漏报. 同时, 它不需要全部的正常和异常的信息, 在给出较少的正常和异常执行迹的情况下就能得到比较理想的检测效果. 由于入侵检测的问题实际训练数据的搜索一般比较困难, 这个特性十分有利于实际系统的应用.

表 1 训练和检测数据的分配情况

	正常执行迹	异常执行迹
训练数据集	260	90
测试数据集	2603	987

表 2 归一化后的三种入侵检测方法的序列异常度性能比较

	Forrest	Lee	本文方法
异常序列最大值	100	100	100
异常序列最小值	5.66	28.68	25.90
正常序列	0	4.41	2.24

4 结束语

通过理论分析和对 Sendmail 系统调用数据集的实验验证表明, 本文提出的基于构造性核覆盖的入侵检测方法, 只需要较少的训练数据, 就能得到较好的结果, 具有较高的检测性能和较快的检测速度, 不但克服了一般机器学习算法收敛速度慢, 易于陷入局部最小点的问题, 而且在一定意义上考虑到系统结构的优化问题(指网络规模最小). 如果将基于此模型的入侵检测算法用于实时检测, 对系统性能的影响较小, 是一种高效、低负荷的检测方法.

参考文献:

[1] Anup K Ghosh. Aaron Schwartzbard. A study in using neural networks for anomaly and misuse detection[A]. The 8th USENIX Security Symposium[C]. Washington D C, 1999.46 – 57.

[2] Balajinath B, Raghavan S V. Intrusion detection through learning behavior model[J]. Computer Communication, 2001, 24(12): 1202– 1212.

[3] 张凤斌, 杨永田, 江子扬. 遗传算法在基于网络异常入侵检测中的应用[J]. 电子学报, 2004, 32(5): 875– 877.
Zhang Feng bin, Yang Yong tian, Jiang Zi yang. Genetic algor

rithms in intrusion detection based on network anomaly[J]. Acta Electronica Sinica, 2004, 32(5): 875- 877. (in Chinese)

[4] Jha S, Tan K, Maxion R A. Markov Chains, classifiers and intrusion detection[A]. The 14th IEEE Computer Security Fourr dations workshop[C]. Canada, 2001. 206- 215.

[5] 张箭, 龚俭. 一种基于模糊综合评判的入侵异常检测方法[J]. 计算机研究与发展, 2003, 40(6): 77- 6782.
Zhang Jian, Gong Jian. An anomaly detection method based on fuzzy judgement[J]. Journal of Computer Research and Development, 2003, 40(6): 77- 6782. (in Chinese)

[6] 饶鲜, 董春曦, 杨绍全. 基于支持向量机的入侵检测系统[J]. 软件学报, 2003, 14(4): 798- 803.
Rao Xian, Dong Churr xi, Yang Shaor quan. An intrusion detection system based on support vector machine[J]. Journal of Software, 2003, 14(4): 798- 803. (in Chinese)

[7] 吴涛, 张铃, 张燕平. 机器学习中的核覆盖算法[J]. 计算机学报, 2005, 28(8): 1295- 1301.
Wu Tao, Zhang Ling, Zhang Yarr ping. Kernel covering algorithm for machine learning[J]. Chinese Journal of Computers, 2005, 28(8): 1295- 1301. (in Chinese)

[8] 张铃, 张钹. M-P 神经元模型的几何意义及其应用[J]. 软件学报, 1998, 9(5): 334- 338.
Zhang Ling, Zhang Bo. A geometrical representation of M-P neural model and its applications. Journal of Software, 1998, 9(5): 334- 338. (in Chinese)

[9] 张学工译. 统计学习理论的本质[M]. 北京: 清华大学出版社, 1995.

[10] Wilson D, Martinez R. Improved heterogeneous distance functions[J]. Journal of Artificial Intelligence Research, 1997, 6(1): 1- 34.

[11] 李辉, 管晓红, 咎鑫, 韩崇昭. 基于支持向量机的网络入侵检测[J]. 计算机研究与发展, 2003, 40(6): 799- 807.
Li Hui, Guan Xiaor hong, Zan Xin, Han Chongr zhao. Network intrusion detection based on support vector machine[J]. Journal of Computer Research and Development, 2003, 40(6): 799- 807. (in Chinese)

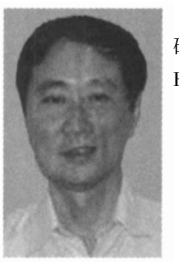
[12] http://www.cs.unm.edu/~immsee/systemcalls.htm[C/OL]. 2001-06-18.

[13] Lee W, Stolfo S J. A data mining framework for building intrusion detection medel[A]. Proceedings of 1999 IEEE Symposium on Security and Privacy[C]. Oakland, CA: IEEE Computer Society Press, 1999. 120- 132.

[14] Forrest S, Hofmeyr S A, et al. A sense of self for unix process[A]. Proceedings of 1996 IEEE Symposium on Computer Security and Privacy[C]. Canada, 1996. 120- 128.

[15] Lee W, Stolfo S, Chan P. Learning patterns from unix process execution traces for intrusion detection[A]. Proceeding of AAAI workshop: AI Approaches to Fraud Detection and Risk Management[C]. Washington D C, 1997. 191- 197.

作者简介:



周鸣争　男, 1958 年生于安徽枞阳, 教授, 研究方向为人工智能、计算机网络与信息安全. E-mail: mzzhou@ah.edu.cn

楚　宁　女, 1980 年生, 硕士研究生, 研究方向为计算机网络安全.

强　俊　女, 1981 年生, 硕士研究生, 研究方向为人工智能与模式识别.