一种基于粗糙集属性约简的支持向量异常入侵检测方法*)

张义荣　鲜　明　肖顺平　王国玉
(国防科技大学电子科学与工程学院　长沙 410073)

摘　要　实现了一种粗糙集属性约简和支持向量机分类相结合的异常入侵检测方法。针对网络连接记录特征属性高维的特点,采用粗糙集属性约简的方法压缩数据空间,然后采用 v-SVM 两分类方法处理约简和正规化后的数据。基于 DARPA 1998 数据源的实验表明,与采用全部属性的 v-SVM 两分类方法相比,该方法具有与之相当的分类精度,但有效地降低了检测时间,减少了存储空间。

关键词　异常检测,粗糙集理论,属性约简,v-SVM 算法,异构值差度量(HVDM)

An Anomaly Intrusion Detection Technique of Support Vector Machine Based on Rough Set Attribute Reduction

ZHANG Yi-Rong　XIAN Ming　XIAO Shun-Ping　WANG Guo-Yu
(School of Electronic Science and Engineering, National Univ. of Defense Technology, Changsha 410073)

Abstract　This paper presented the implementation of a hybird anomaly intrusion detection technique based on rough set attribute reduction and support vector machine(SVM). According to the high dimension of network records with feature attributes, the rough set attribute reduction approach is firstly utilized to reducing data space and then the v-SVM algorithm is introduced into processing normalized data set. Experiments on DARPA 1998 data set show that the proposed anomaly detection technique achieves a comparable precise detection rate as the v-SVM algorithm based on all feature attributes, however, evidently decreases detection time as well as storage space.

Keywords　Anomaly detection, Rough set theory, Attribute reduction, v-SVM algorithm, Heterogeneous value difference metric(HVDM)

1　引言

入侵检测(Intrusion Detection)是指对于面向计算机和网络资源的恶意行为的识别和响应。根据 IDS 分析引擎中使用的检测方法的不同,可以把 IDS 分为误用检测和异常检测。所谓误用检测是指运用已知攻击方法,根据已定义好的入侵模式,通过判断这些入侵模式是否出现来检测。而异常检测是指根据使用者的行为或资源使用状况的正常程度来判断是否入侵。由于异常检测可以发现针对系统的未知网络攻击,因此是入侵检测研究的热点。目前人们主要采用了数据挖掘[1]、隐马尔可夫模型(HMM)[2]、遗传算法[3]、人工免疫系统[4]等机器学习方法来建立系统运行的正常轮廓或正常行为与异常行为的差异,从而达到提高检测率、降低误警率的目的。然而上述机器学习方法都具有统计"渐进性",即学习器的性能跟训练数据量的大小有密切关系,在学习过程中往往需要系统提供大量的正常行为数据,并且要求数据具有一定的正规性。在训练样本数不足时,系统性能有明显的下降。然而,在入侵检测中要获得大量的纯净的训练数据并不容易,实际获得的数据源具有多变性、不同质、高维数、小样本等特性。

统计学习理论(SLT)[5]克服了传统的统计学中要求样本趋于无穷的要求,它给出了小样本情况下学习风险与样本分布无关的界,并在结构风险最小化(SRM)准则之下提供了使界达到最小的一种方法-支持向量机(SVM)。SVM 和核学习方法的解的稀疏性、对样本维数的不敏感性和良好的分类精度使得 SVM 在模式识别中得到了广泛的应用,如手写识别、人脸识别、文本分类等。从统计学习理论的角度来看,入侵检测可被视为一种模式识别中的分类问题,即根据网络流量特征(目的地址、源地址、目的端口号、源端口号、传输协议、发送字节数、TCP 选项等)和主机审计记录(CPU 利用率、I/O 利用率、文件访问、用户命令调用序列)等区分系统的正常行为和异常行为。在把 SVM 技术引入到入侵检测中,人们已做了不少工作[6~8]。

基于 SVM 技术的入侵检测问题通常转化为一个二次规划问题来求解。但二次规划的计算量随着变量的增加而呈指数增加。对于在入侵检测中经常遇到的高维、大规模数据的模式分类问题,如何提高基于 SVM 进行数据处理的实时性、缩短训练和检测时间,是当前一个需要解决的重要问题。为了适应实时异常检测的要求,有必要在 SVM 训练之前进行样本属性的特征选择,以降低 SVM 分类器的复杂度,提高检测速度。粗糙集理论作为一种较新的机器学习方法,在处理不确定的知识、消除冗余信息、发现样本数据属性之间的本质关系上具有突出的优势,它不依赖模型的先验知识,提供了一套完整的条件属性约简和值约简方法,从而可以找到描述系统正常模型的最小预测规则集,为完成特征属性选择和提高检测速度提供了新的途径。

*)本课题得到国家自然科学基金项目(60372039)和"十五"国防预研基金项目(41329080101)资助。张义荣　博士研究生,主要研究方向为信息安全和智能信息处理;鲜　明　博士、副教授,主要研究方向为信息安全与电子对抗;肖顺平　教授、博士生导师,主要研究方向为电子系统建模、仿真与评估;王国玉　研究员、博士生导师,主要研究方向为信息对抗与目标识别。

本文实现了一种粗糙集属性约简和支持向量机分类相结合的网络异常入侵检测方法。针对网络流量记录特征属性高维的特点，采用粗糙集属性约简的方法压缩数据空间，在不损失任何有效信息的前提下降低了数据处理的规模，然后采用 v-SVM 两分类方法处理约简和正规化后的数据。基于 DARPA 1998 数据源[7]的实验表明，与基于全部属性的 v-SVM 两分类方法相比，该方法具有与之相当的分类精度，但有效地降低了分类时间，减少了存储空间。

2 粗集知识约简概述[9]

2.1 决策表系统与粗糙集

定义 1 信息表知识表达系统 S 是一个四元组 $S=\langle U,R,V,f\rangle$，其中 U 是一个非空的有限对象集，称为论域；$R=C\cup D$ 属性集合，子集 C 和 D 分别称为条件属性集和结果属性集；$V=\bigcup_{r\in R}V_r$ 是属性值的集合，V_r 表示属性 $r\in R$ 的属性值范围；$f:U\times R\rightarrow V$ 是一个信息函数，它指定 U 中每一个对象 x 的属性值。

一个决策表是一个信息表知识表达系统 $S=\langle U,R,V,f\rangle$，$R=C\cup D$ 属性集合，子集 C 和 D 分别称为条件属性集和结果属性集，$D\neq\varnothing$。条件属性 C 和结果属性 D 的等价关系 $IND(C)$ 和 $IND(D)$ 的等价类分别称为条件类和决策类。

定义 2 令 $X\subseteq U$，当 X 能用属性子集 B 确切地描述(即是属性子集 B 所确定的 U 上的不分明集的并)时，称 X 是 B 可定义的。B 可定义集也称作 B 精确集，B 不可定义集称为 B 粗糙(Rough)集。

定义 3 对每个概念 X 和不分明关系 B，包含于 X 中的最大可定义集和包含 X 的最小可定义集，分别称为 X 的下近似集($B_-(X)$)和上近似集($B^-(X)$)。显然

$$B_-(X)\subseteq X\subseteq B^-(X)$$

根据定义，

$$B_-(X)=\{x\mid x\in U\wedge[x]_B\subseteq X\}$$

$$B^-(X)=\{x\mid x\in U\wedge([x]_B\cap X\neq\varnothing)\}$$

集合 $BN_B(X)=B^-(X)\backslash B_-(X)$ 称为 X 的 B 边界；$POS_B(X)=B_-(X)$ 称为 X 的 B 正域；$NEG_B(X)=U\backslash B_-(X)$ 称为 X 的 B 负域。

定义 4 假定集合 X 是论域 U 上的一个关于知识 B 的 Rough 集，定义其 B 粗糙度 $P_B(X)$ 为：

$$P_B(X)=1-|B_-(X)|/|B^-(X)|$$

其中 $X\neq\varnothing$。由此可见，$P_B(X)$ 为一个区间[0,1]上的实数，它定义了集合 X 的确定度。

定义 5 令决策表系统 $S=\langle U,R,V,f\rangle$，$R=P\cup D$ 为属性集合，子集 $P=\{a_i\mid i=1,\cdots,m\}$ 和 $D=\{d\}$ 分别称为条件属性集和决策属性集，$U=\{x_1,x_2,\cdots,x_n\}$ 是论域，$a_i(x_j)$ 是样本 x_j 在属性 a_i 上的取值。$C_D(i,j)$ 表示可辨识矩阵中第 i 行、第 j 列的元素，则可辨识矩阵 C_D 定义为

$$C_D(i,j)=\begin{cases}\{a_k\mid a_k\in P\wedge a_k(x_i)\neq a_k(x_j)\} & d(x_i)\neq d(x_j)\\ 0 & d(x_i)=d(x_j)\end{cases}$$

2.2 决策表离散化

在运用 Rough 集理论处理决策表时，要求决策表中的值用离散数据表达。如果某些条件属性或决策属性的值域为连续值，则在处理前必须进行离散化处理。常用的离散化处理方法有等距离划分法、等频率划分法、Naïve Scaler 算法、Semi Naïve Scaler 算法、Boolean 逻辑和 Rough 集理论相结合的离散化算法及基于属性重要性的离散化算法等。如对常用的 Naïve Scaler 算法，处理过程如下：

对每一属性 $a\in C$，进行下面的过程：

- 根据 $a(x)$ 的值，由小到大排列实例 $x\in U$；
- 从上到下扫描，设 x_i 和 x_j 代表两个相邻的实例：若 $a(x_i)=a(x_j)$，则继续扫描；否则，若 $d(x_i)=d(x_j)$，则决策相同，继续扫描；否则，得到一个断点 c，$c=(a(x_i)+a(x_j))/2$。

2.3 决策表属性约简

属性约简就是要从条件属性集合中发现部分必要的条件属性，使得根据这部分条件属性形成的相对于决策属性的分类和所有条件属性形成的相对于决策属性的分类一致，即和所有条件属性相对于决策属性 D 有相同的分类能力。

定义 6 设 U 为一个论域，P 和 Q 为定义在 U 上的两个等价关系簇，Q 的 P 正域记为 $POS_P(Q)$，定义为：

$$POS_P(Q)=\bigcup_{X\in U/Q}P_-(X)$$

进一步，$POS_P(Q)=POS_{(P\backslash\{r\})}(Q)$，则称 r 为 P 中相对于 Q 可省略的。若 P 中每一 r 都是 P 中相对于 Q 不可省略的，则称 P 相对于 Q 独立。

定义 7 设 U 为一个论域，P 和 Q 为定义在 U 上的两个等价关系簇，若 P 的 Q 独立子集 $S\subset P$，满足 $POS_S(Q)=POS_P(Q)$，则称 S 为 P 的 Q 约简，记为 $RED_Q(P)$。P 的所有 Q 约简的交称为 P 的 Q 核，记为 $CORE_Q(P)$。

决策表属性约简为一 NP-hard 问题，相应的算法有一般约简算法、基于可辨识矩阵和逻辑运算的约简算法、归纳属性约简算法、基于互信息的属性约简算法和基于特征选择的属性约简算法等。对于基于可辨识矩阵和逻辑运算的约简算法，过程如下：

Step1：计算决策表的可辨识矩阵 C_D；

Step2：计算决策表的可辨识矩阵中的所有取值为非空集合的元素 C_{ij} ($C_{ij}\neq 0, C_{ij}\neq\varnothing$)，建立相应的析取逻辑表达式 L_{ij}，$L_{ij}=\bigvee_{a_i\in C_{ij}}a_i$；

Step3：将所有的析取逻辑表达式 L_{ij} 进行合取运算，得一个合取范式 L；

$$L=\bigwedge_{C_{ij}\neq 0,C_{ij}\neq\varnothing}L_{ij}$$

Step4：将合取范式 L 转化为析取范式的形式，得：

$$L'=\bigvee_i L_i$$

Step5：输出属性约简结果。析取范式中的每个合取项就对应一个属性约简的结果，每个合取项中包含的属性组成约简后的条件属性集合。

3 v-SVM 分类基本原理[10]

对于两类的模式识别问题，假设有 l 个独立同分布的观测样本：

$$\{(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l)\}\subset R^N\times\{\pm 1\} \quad (1)$$

假设样本是线性可分的，根据统计学习理论，为了使泛化误差最小，我们需要找到 VC 维最小的间隔超平面，即最大间隔分类超平面。在样本线性不可分的情况下，需要引入松弛变量 ξ_i，并使得分类超平面的范数和与错误惩罚项之和达到最小，如图 1 所示。在 v-SVM 中引入参数 v 和变量 ρ，最小化下式：

$$\tau(w,\xi,\rho)=\frac{1}{2}\|w\|^2-v\rho+\frac{1}{l}\sum_i\xi_i \quad (2)$$

满足约束：

$$y_i((w \cdot x_i)+b) \geqslant \rho-\xi_i \tag{3}$$

$$\xi_i \geqslant 0, \rho \geqslant 0 \tag{4}$$

上述最优化问题的对偶形式为最小化下式：

$$W(a)=\frac{1}{2}\sum_{i,j}a_i a_j y_i y_j (x_i \cdot x_j) \tag{5}$$

满足约束

$$\sum_i a_i \geqslant v \tag{6}$$

$$0 \leqslant a_i \leqslant \frac{1}{l} \tag{7}$$

$$\sum_i a_i y_i = 0 \tag{8}$$

决策函数 $f(x)$ 为

$$f(x)=\mathrm{sgn}(\sum_i a_i y_i (x_i \cdot x)+b) \tag{9}$$

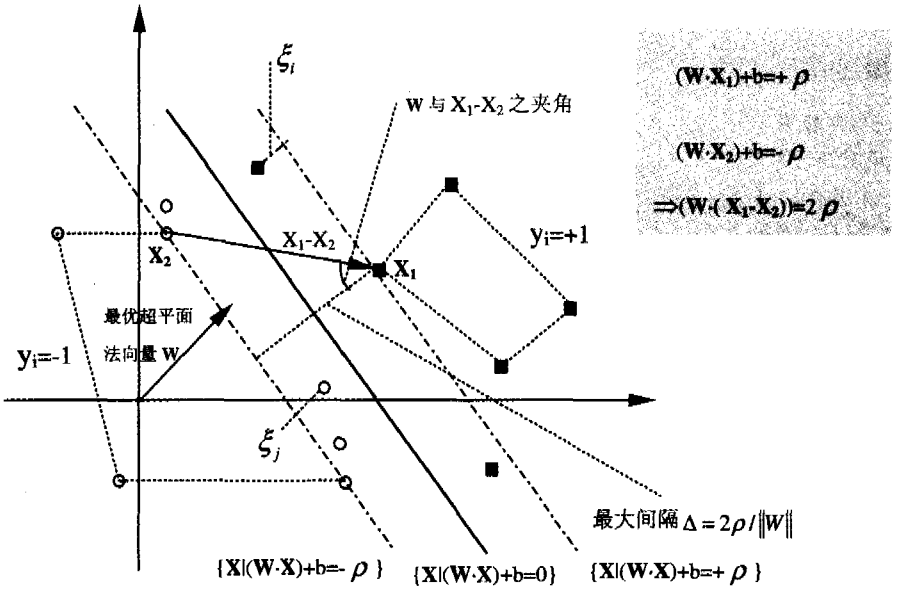

图 1　v-SVM 示意图

在非线性条件下，引入 Mercer 核函数，式(5)转化为

$$W(a)=\frac{1}{2}\sum_{i,j}a_i a_j y_i y_j k(x_i \cdot x_j) \tag{10}$$

相应地，决策函数 $f(x)$ 为

$$f(x)=\mathrm{sgn}(\sum_i a_i y_i k(x_i \cdot x)+b) \tag{11}$$

上述对偶问题为一二次规划问题，其解具有全局最优性。根据 KKT 条件，在数据集中只有部分数据使得式(3)取等号，这样的样本点称之为支持向量。通过调整参数 v 可以控制 SVM 的泛化能力，根据 B. Schölkopf 等人[10]的结论，参数 v 是训练样本集中孤立点所占比例的上界，也是训练样本集中支持向量所占比例的下界。

这样，输入空间的两分类问题就转化为二次规划问题，这个规划问题有相应的快速解法。采用不同的满足 Mercer 条件的函数作为核函数，就可以构造实现输入空间中不同类型的非线性决策面的学习机器。

4　基于粗糙集属性约简的 v-SVM 异常入侵检测系统

基于粗糙集属性约简的 v-SVM 异常入侵检测系统的基本思想是利用粗糙集理论对输入空间的维数进行约简，完成输入特征提取工作，从而达到降低数据处理的规模的目的。根据约简得到的最小条件属性集及相应的原始数据重新形成新的训练样本集，该样本集仅保留了影响分类精度的重要属性。约简后形成的训练样本进行数值化和归一化，然后输入 v-SVM 学习器训练。最后输入按照最小条件属性集及相应的原始数据形成的测试样本集对系统进行测试，并输出分类结果。整个过程如图 2 所示。

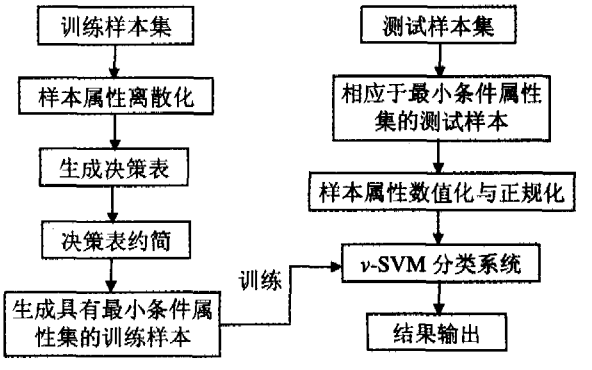

图 2　基于粗糙集属性约简的 v-SVM 异常入侵检测过程

5　实验结果分析

实验中的数据源来自 1998 年由 MIT 的 Lincoln 实验室 Wenke Lee[11] 等人为美国国防部高级研究计划署(DARPA)负责实施的入侵检测评估项目，他们仿真了空军的一个典型的局域网。这个数据源由网络数据流量、主机审计记录和系统文件转储(dump)得到，数据源包含了 7 周的训练数据(压缩为 4GB)和 2 周的测试数据，大约有 500 万条训练数据和 30 万条测试数据。测试数据包含了 38 种攻击，分为 4 大类：Probing、Denial of Service(DoS)、User-to-Root(U2R)和 Remote-to-Local(R2L)，其中 14 种攻击在训练数据中没有出现过。

一个完整的 TCP 会话被认为是一个连接记录，每条连接信息由 4 类属性集组成：基本属性集、(基于时间的)流量属性集、基于主机的流量属性集和内容属性集，共包含 41 个不同性质的属性，其中含有 32 个连续性属性和 9 个离散属性。这

9个离散属性是protocol_type、service、flag、land、logged_in、root_shell、su、hot_login和guest_login。Wenke Lee等人发现，对于不同类型的攻击，需要采用不同的特征属性子集。对于U2R和R2L类攻击，主要采用了基本属性集和内容属性集；而对于Probing和DoS类攻击，则需采用基本属性集、(基于时间的)流量属性集和基于主机的流量属性集。

5.1 TCP记录特征属性约简

由于原始数据集过于庞大，这里选用了一个10%的数据集Correct作为实验数据，它包含正常记录和4类攻击记录。攻击记录中各类攻击所占的比例为：DoS,93%；Probing,2%；U2R,2%；R2L,2%。一个TCP记录的41个特征属性中含有32个连续性属性，采用Naïve Scaler算法离散化，这样所有的特征属性全部转化为离散属性。条件属性的约简采用波兰华沙大学与挪威科技大学联合开发的Rosetta[12]软件实现。实验结果如下：

- 正常记录：约简属性集(26个)为{1,3,5-7,8-10,14,15,17,20-23,25-29,33,35,36,38,39,40}；
- DoS：约简属性集(18个)为{1,3,5,6,23-28,32,33,35,36,38-41}；
- Probing：约简属性集(7个)为{3,5,6,23,24,32,33}；
- U2R：约简属性集(8个)为{5,6,8,15,16,18,32,33}；
- R2L：约简属性集(8个)为{3,5,6,21,22,24,32,33}。

上述结果与文[11]基本相符，与文[13]的结果有少许差别，这可能是决策表离散化过程中的差异和训练样本集选择的差异引起的。

5.2 约简属性的数值化

经Rough集理论约简后的结果大大缩小了网络连接记录的维数，从而有利于提高v-SVM算法的分类精度。由于v-SVM只能处理数值化向量，因此在把训练样本输入v-SVM学习器前，必须对特征属性进行预处理。这里采用了和文[14]一样的处理方法，即引入异构值差度量(HVDM)距离函数，来反映不同属性对异构数据集样本点间距离的贡献。假设异构数据集X上两个数据x,y的第i个连续属性分别为x_i,y_i，则x,y第在i个属性上的距离为：

$$normalized\text{-}diff_i(x,y)=\frac{|x_i-y_i|}{4\sigma_i} \tag{12}$$

其中，σ_i为数据集上第i个属性的方差。

假设异构数据集X上两个数据x,y的第a个离散属性分别为x_a,y_a，则x,y在第a个属性上的值差度量为：

$$normalized\text{-}vdm_a(x,y)=\sqrt{\sum_{c=1}^{C}\left|\frac{N_{a,x,c}}{N_{a,x}}-\frac{N_{a,y,c}}{N_{a,y}}\right|^2} \tag{13}$$

其中，$N_{a,x}$为数据集X上所有数据第a个属性取值为x_a的数据的个数，$N_{a,x,c}$为数据集X上所有数据第a个属性取值为x_a且输出类别为C的数据的个数，C为数据集的所有输出类别。

异构值差度量(HVDM)距离函数$H(x,y)$定义为：

$$H(x,y)=\sqrt{\sum_{a=1}^{m}d_a^2(x_a,y_a)} \tag{14}$$

式(14)中，$d_a(x_a,y_a)$的取值如下：若x_a或y_a未知，取值为1；若x_a和y_a为离散属性，由式(13)计算；若x_a和y_a为连续属性，由式(12)计算。

5.3 基于约简结果的v-SVM异常检测结果

根据上述特征属性约简结果，构造了4个训练样本集和4个测试样本集。训练数据集1(TrS1)由网络正常会话和DoS攻击记录组成，共有记录15640个，其中正常会话8800个，是从原始训练数据集中随机抽取出来的，包含各种协议(TCP、UDP、ICMP)和各种服务(http、ftp、telnet等)，DoS攻击记录6840个，由DARPA 1998的训练子集的DoS攻击记录中随机抽取得到；测试数据集1(TeS1)共有记录13764个，其中正常会话7560个，是从原始测试数据集中随机抽取出来的，而DoS攻击记录6204个，来自于DARPA 1998的测试子集。

训练数据集2(TrS2)由网络正常会话和Probing攻击记录组成，共有记录12640个，其中正常会话7568个，Probing攻击记录5072个；测试数据集2(TeS2)共有记录11764个，其中正常会话6560个，而Probing攻击记录5204个，来自于DARPA 1998的测试子集。

训练数据集3(TrS3)由网络正常会话和U2R攻击记录组成，共有记录6548个，其中正常会话3948个，U2R攻击记录2600个，由原始训练数据集中全部52条U2R攻击记录复制50份得到；测试数据集3(TeS3)共有记录5764个，其中正常会话3484个，而U2R攻击记录2280个，由DARPA 1998的测试子集中全部228条U2R攻击记录复制10份得到。

训练数据集4(TrS4)由网络正常会话和R2L攻击记录组成，共有记录6653个，其中正常会话4401个，R2L攻击记录2252个，由原始训练数据集中全部1126条U2R攻击记录复制2份得到；测试数据集4(TeS4)共有记录8749个，其中正常会话5348个，而R2L攻击记录3401个，来自于DARPA 1998的测试子集。

v-SVM的核函数取为：

$$K(x_i,x_j)=\exp(-H(x_i,x_j)/2\sigma^2)$$

其中$H(x_i,x_j)$的定义由式(14)给出。v-SVM分类器的性能与参数v和σ^2有关，通常v越小，σ^2越大，则解所包含的支持向量越少，学习器的泛化性能越好。这里取v为0.1，σ^2为20，训练和检测过程由LibSVM软件包[16]实现，实验结果如表1所示。

表1 基于粗糙集属性约简的v-SVM异常检测结果

类别	训练时间(s)	检测时间(s)	检测精度(%)
正常	12.3	1.86	99.13
DoS	26.4	1.97	99.34
Probing	31.1	2.13	99.56
U2R	4.38	0.79	98.82
R2L	9.58	0.64	98.69

作为对比，利用上述训练和测试数据集，在不做特征属性约简的情况下，使用同样的参数，由v-SVM学习器得到的检测性能如表2所示。

表2 基于全部特征属性的v-SVM异常检测结果

类别	训练时间(s)	检测时间(s)	检测精度(%)
正常	11.3	1.94	99.12
DoS	29.4	2.21	99.44
Probing	39.7	2.45	99.51
U2R	6.52	1.06	99.12
R2L	13.29	0.94	99.06

由表1和表2的结果可知，基于粗糙集属性约简的v-SVM异常检测与基于全部41个特征属性的v-SVM异常检测结果相比，两者具有相当的检测精度，且都保持着很高的检

测率，前者在训练和检测时间上有所缩短，对于有的攻击类型(U2R、R2L)的检测，检测速度提高得比较显著。注意到U2R和R2L攻击记录在整个数据集中数量较少，为了提高υ-SVM异常检测对U2R和R2L攻击检测的性能，在构造训练和测试集时，我们对U2R和R2L攻击记录做了多次复制，以混进正常样本集，这样可以得到比较均衡的数据集。

结论 在基于SVM的异常入侵检测中，提高检测速度对于实时入侵检测是十分重要的。提高检测速度有几种途径：一是对硬件设备进行升级，如增大存储器的容量、提高CPU的处理速度；一是研究二次规划的快速算法；再一个是对网络连接记录这种高维数据进行特征选择和降维，以减小数据处理的规模。本文着眼于第三种提高检测速度的方法，实现了一种粗糙集属性约简和支持向量机分类相结合的网络异常入侵检测方法。首先利用粗糙集属性约简的方法压缩数据空间，然后采用υ-SVM两分类方法处理约简和正规化后的数据。实验结果表明，与基于全部属性的υ-SVM两分类方法相比，该方法具有与之相当的分类精度，但有效地缩短了分类时间，减少了存储空间。

参考文献

1 Lee W, Stolfo S J. Data mining approaches for intrusion detection [A]. In: Proceedings, Seventh USENIX Security Symposium, San Antonio, TX, 1998

2 Jha S, Tan K, Maxion R. Markov chains, classifiers and intrusion detection [A]. In: The 14th IEEE Computer Security Foundations Workshop, Canada, 2001. Proceedings, Seventh USENIX Security Symposium, San Antonio, TX, 1998

3 Balajinath B, Raghavan S. Intrusion detection through learning behavior model [J]. Computer Communications, 2001, 24(12): 1202～1212

4 Forrest S, Hofmeyr S A. Computer Immunology [J]. Communications of the ACM, 1997, 40(10): 88～96

5 Vapnik V N. The nature of statistical learning theory [M]. New York: Springer, 1995

6 Kim D S, Park J S. Network-based intrusion detection with support vector machines [A]. In: KahngH-K. Ed. ICOIN 2003, LNCS 2662, 2003. 747～756

7 Sohn T, Seo J T, Moon J S. A study on the covert channel detection of TCP/IP header using support vector machines [A]. In: Qing S, Gollmann D, Zhou J. Eds. ICOIN 2003, LNCS 2836, 2003. 313～324

8 Hu W J, Liao Y H, Vemuri V R. Robust anomaly detection using support vector machines [J]. IEEE Transactrions on Pattern Analysis and Machine Intelligence. (in press)

9 王国胤. Rough集理论与知识获取[M]. 西安：西安交通大学出版社，2001

10 Schölkopf B, Smola A, Williamson R C, et al. New support vector algorithms. Neural Computation [J]. 2000, 12(5): 1207～1245

11 Lee W, Stolfo S. A framework for constructing features and models for intrusion detection systems [J]. ACM Transactions on Information and System Security, 2000(3): 227～261

12 http://rosetta.lcb.uu.se/general/

13 Sung A H, Mukkamala S. Identifying Important Features for Intrusion Detection Using Support Vector Machines and Neural Networks [A]. In: Proceedings of 2003 Symposium on Application and the Internet(SAINT'03), 2003

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

15 李元诚，方廷健. 一种基于粗糙集理论的SVM短期负荷预测方法[J]. 系统工程与电子技术，2004, 26(2): 187～190

16 Chang Chih-Chung, Lin Chih-Jen. LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/～cjlin/libsvm

(上接第63页)

别着不同的颜色。当某个存储节点上有存储对象要迁移时，其相应的有色托肯将激活变迁T_{10}，请求元数据服务器判别如何调度，同时根据当前所有存储节点的对象负载情况，确定目标OSD，并实现对象的迁移。

结束语 对象存储系统是下一代互联网存储模式的关键技术，而对象存储系统中对象管理，其成功与否在一定的程度上关系着整个存储系统的成功建立与否。网络对象存储系统中存储节点的动态变化是一项很重要的特点，这样的特点决定了对象存储系统中对象的迁移是经常性的，而且也是必须的。因为存储对象原来的节点暂时不存在了，但是整个系统不能因为存储节点的暂时不在而不能正常运行。为了解决这样的问题，对象迁移是一种非常好的解决办法。本文对对象迁移的策略进行了理论上的分析，通过可变阈值和阈长的动态反馈调整模型确定对象存储系统中目标OSD的选择依据，并引入面向Petri网的相关技术，对对象迁移进行建模和分析，对存储对象的迁移实现了有效的控制。

参考文献

1 Kangasharju J, Roberts J, Ross K W. Object replication strategies in content distribution networks. Computer Communications, 2002, 25(4): 376～383

2 http://www.intel.com/labs/storage/osd/tech.htm, 2003-03

3 Braam P J. The Lustre Storage architecture [EB/OL]. http://www.lustre.org/docs/lustre.pdf, 2003-03

4 Laoutaris N, Zissimopoulos V, Stavrakakis I. Joint object placement and node dimensioning for internet content distribution. Information Processing Letters, 2004, 89(6): 273～279

5 Krishnan P, Raz D, Shavit Y. The cache location problem. IEEE/ACM Transactions on Networking, 2000, 8(5): 568～581

6 Ranganathan K, Foster I. Identifying dynamic replication strategies for a high performance data grid. Proceedings of the International Workshop on Grid Computing, Denver, Colorado. 2001

7 Bowden F D J. Modeling time in PetriNets. The Second Australia, Japan Workshop on Stochastic Models, Gold Coast, 1996

8 Zhou Long-xiang. C2POREL22: A distribute drelational database management system on microcomputer network. Science, Series A, 1986, XXIX(1): 78～91

9 Tlin J. A Petrinet based integrated control and scheduling scheme for flexible manufacturing cells [J]. Computer integrated Manufacturing System, 1997, 10(2): 109～122

10 Wu C H, Lee S J. Enhanced high-level Petri nets with multiple colors and knowledge verification/validation of rule-based expert system [J]. IEEE Trans on System, Man, and Cybernetics, 1997, 27(5): 73～85

11 Montague R M, Denegri S L, Curlin T H, et al. System Area Networks: Next Generation of Scale in the Data Center [Z]. Dain Rauscher Incorporated, 2001

12 袁崇义. Petri网原理. 北京：电子工业出版社，1998

13 杜兴，谢立，孙中秀. 一种基于对象的分布式系统描述求精方法. 计算机学报，1994, 17(7): 562～567

14 丁志军，蒋昌俊. 并发程序验证的时序Petri网方法[J]. 计算机学报，2002, 25(5): 62～69