第 34 卷 第 4 期 计 算 机 工 程 2008 年 2 月
Vol.34 No.4 Computer Engineering February 2008
·安全技术· 文章编号：1000—3428(2008)04—0154—03 文献标识码：A 中图分类号：TP274.4

基于 SVM 的数据融合方法在 DIDS 中的应用

叶 苗[1,2]，王 勇[1]，麦范金[2]，陈超泉[2]
(1. 桂林电子科技大学网络中心，桂林 541004；2. 桂林工学院网络中心，桂林 541004)

摘 要：考虑到传统 SVM 解决传统 IDS 问题的困难，建立基于带概率输出信息的 SVM 局部信息检测和数据融合、决策分析的分布式入侵检测 DIDS 模型。该模型尽可能利用局部 SVM 分类器的优势，充分考虑了各局部 SVM 的性能差别。通过 KDD99 数据集对该模型的测试，证明该分布式入侵检测模型可以明显地降低入侵检测的漏报率，提高检测精度。

关键词：支持向量机；概率分配函数；分布式入侵检测；数据融合；检测率

Application of SVM Sensor and Data Fusion in Distributed Intrusion Detection System

YE Miao[1,2], WANG Yong[1], MAI Fan-jin[2], CHENG Chao-quan[2]
(1. Network Center, Guilin University of Electronic Technology, Guilin 541004;
2. Network Center, Guilin Institute of Industrial Technology, Guilin 541004)

【Abstract】To solve the difficulty of traditional SVM applied into IDS, a distributed intrusion detection model based on SVM sensor with probability estimation and data fusion is proposed. The local SVM's advantage and differences among each local SVM's performance are considered in this model. Experimental results carried out with KDD99 dataset show that the model can make false positive lower and improve the efficiency of the intrusion detection

【Key words】support vector machine; probability assignment function; distributed intrusion detection; data fusion; detection rate

传统的入侵检测系统，主要以集中式控制和基于模式匹配的误用检测系统为主[1-3]。模式匹配技术对已知攻击类型的攻击检测率高，但对未知的攻击检测率低；而集中式的构架自适应能力差，容易遭受攻击和。因此，采用数据挖掘技术的异常检测和分布式构架的IDS成为目前该领域讨论的热点[4-5]。在用于异常检测的数据挖掘技术中，支持向量机(SVM)在小样本、非线性及高维模式识别问题的处理方面具有独特的优势[5]，适合处理入侵检测领域中的高维异构不均衡数据集。具有融合中心的分布式IDS可以解决传统IDS适应性和扩展性差的问题。

1 支持向量机(SVM)及其概率意义的推广

1.1 支持向量机简介

标准 SVM 的决策函数是

$$y = f(x) = \sum_{i=1}^{n} y_i \alpha_i K(x, x_i) + b \tag{1}$$

其中，$x \in R^d$ 是 d 维矢量；$y \in \{-1,1\}$ 是分类标签；n 是训练集的大小；$K(x, x_i)$ 是核函数；其余的量为参数。求解 SVM 决策函数的参数可以通过求解对偶问题得到，即

$$L(\alpha) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j K(X_i, X_j) - \sum_{j=1}^{N} \alpha_j \tag{2}$$

$$\text{s.t.} \quad \sum_{i=1}^{N} y_i \alpha_i = 0 \tag{3}$$

最大化式(2)得到解参数 $a = \alpha^* = (\alpha_1^*, \alpha_2^*, \cdots, \alpha_n^*)$，对非支持向量数据点对应的 α_i，取值为 0。

1.2 具有概率估计信息的 SVM 简介

SVM用于预测的分类结果是两种极端的情况，即要么属于某一类的概率为 1，要么属于另一类的概率为 1。近来有人提出了具有概率信息分类结果的SVM[6-9]。

用得比较多的是用Simoidal函数作为概率分配函数，即 $P(y=1/f) = \frac{1}{1+e^{Af+B}}$，其中，参数$A$和$B$可以利用最大似然法根据训练集$(f(x),y)$进行估计得到的。在LIBSVM Version 2.82 released 软件包中采用的就是这样的函数形式，后面的实验采用的就是这个软件包和这种形式的分配函数[7-9]。

2 SVM 分类器在分布式入侵检测中的应用

实际检测环境中的数据可以来自不同网段的数据源，即使是同一个网段，也可能会有不同类型的攻击数据同时存在。单点位置的 SVM 分类器判断网络的异常情况，训练的难度大(表现在 SVM 的训练时间长，数据处理存储量大)，更新麻烦，也缺乏数据的一些细节信息。由此构建出采用 SVM 检测技术的分布式入侵检测系统，其框架结构如图 1 所示。

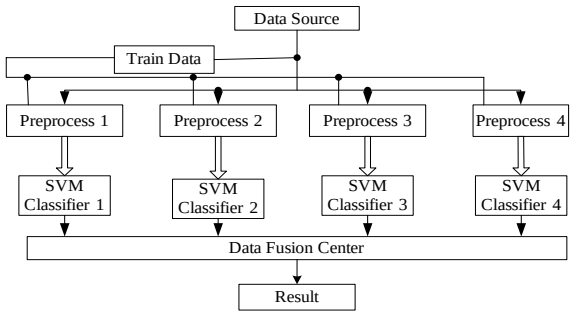

图 1 基于SVM检测技术的分布式入侵检测系统的框架结构

基金项目：广西自然科学基金资助项目(桂科基 0575094)
作者简介：叶 苗(1977－)，男，硕士研究生，主研方向：网络安全，数据挖掘；王 勇，教授、博士；麦范金、陈超泉，副教授
收稿日期：2007-02-28 **E-mail：**acesohn@glite.edu.cn

在图 1 中，局部 SVM 分类器的数据可以是不同信息来源的数据(比如主机型的数据、网络流量方面的数据)，也可以是针对性质类别的数据(比如各种不同攻击类型的异常数据和正常访问的数据)。本文的模型实验中，就是用针对不同攻击类型的异常数据构建局部带概率估计的 SVM 分类器。

决策分类中心进行局部检测结果的数据融合，输出决策结果，判断是否发生异常。目前关于这方面的数据融合的算法有很多[10-12]，可以是逻辑AND或OR运算算法，也可以是加权平均，然后将结果与阈值比较[10]，还可用Bayesian标　准[11]及信息融合领域中的D-S证据理论[12-13]等融合算法。

3　模型构建及融合规则算法

采用 KDD99 数据集来构建一个具体模型，并进行测试。KDD99 数据集[9]包括了大量的网络环境中模拟取得的入侵连接记录。除了正常连接记录外，入侵数据包括 4 大类：DoS(Denial of Service)拒绝/分布式拒绝服务攻击；U2R(User to Root)权限提升攻击；R2L(Remote to Local)远程登录攻击；PRB(probing)端口侦测和扫描攻击。采用针对 4 类攻击分别构建局部SVM分类器，分别检测 4 种类型的攻击，每个分类器分别用来检测正常和异常两种情况。将KDD99 的 1/10 子集(共包括 494 021 条数据)的 4 类入侵连接记录集 TD_{DOS}，$TD_{\mathrm{U2R}}, TD_{\mathrm{R2L}}, TD_{\mathrm{PRB}}$ 和正常连接记录 TD_{NOMAL} 分别提取出来，构建 4 个子集：

$$TD_{\mathrm{D}} = TD_{\mathrm{DOS}} \cup TD_{\mathrm{NOMAL}}, TD_{\mathrm{U}} = TD_{\mathrm{U2R}} \cup TD_{\mathrm{NOMAL}}$$

$$TD_{\mathrm{R}} = TD_{\mathrm{R2L}} \cup TD_{\mathrm{NOMAL}}, TD_{\mathrm{P}} = TD_{\mathrm{PRB}} \cup TD_{\mathrm{NOMAL}}$$

分别作为每个 SVM 分类器的训练集进行训练，得到 4 个 SVM 的决策函数

$$u_1 = f_{\mathrm{DOS}}(x), u_2 = f_{\mathrm{U2R}}(x), u_3 = f_{\mathrm{R2L}}(x), u_4 = f_{\mathrm{PRB}}(x)$$

由于其中采取的是 Simoidal 函数作为概率分配函数值的 SVM 训练机，各个输出值是代表属于正常类+1 类或异常类-1 类的概率值，假设取属于正常类+1 类的概率值为 u 值，则属于-1 类的概率值是 1-u。规定统一将局部检测为正常类+1 类的概率值即将 $u_1, u_2, u_3, u_4 \in [0,1]$ 送到融合中心，如图 2。这样得到融合中心的一个输入向量 $\boldsymbol{U} = [u_1, u_2, u_3, u_4]$。

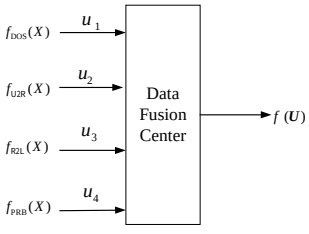

图 2　融合中心

为了构建和检验决策中心的融合算法规则，对KDD 99 的另一个数据子集——带有正确分类标签的子集[9](为了后面叙述方便，称之为KDD99 标签集，包含 29 2241 条连接记录)做类似的处理，分别抽取其异常子集 $TD_{\mathrm{DOS}}, TD_{\mathrm{U2R}}, TD_{\mathrm{R2L}}, TD_{\mathrm{PRB}}$ 和正常子集 TD_{NOMAL} 分别进行测试，得到的正确分类的检测率统计如表 1 所示。从表中可以看出，异常子集 $TD_{\mathrm{DOS}}, TD_{\mathrm{U2R}}, TD_{\mathrm{R2L}}, TD_{\mathrm{PRB}}$ 的检测率相差比较大。运用统计数学理论可以证明，SVM的检测精度与样本的结构有关系[14]。样本数越少，正确的检测率越低；正负样本的比例相差越大，正确检测的检测率越低。从表 2 和表 3 给出KDD 99 的 1/10 子集的各个数据异常子集 $TD_{\mathrm{D}}, TD_{\mathrm{U}}, TD_{\mathrm{R}}, TD_{\mathrm{P}}$ 的大小和各个SVM分类器的正、负支持向量的个数值。显然，各局部SVM分类器的样本分布很不均衡。

表 1　正确分类的检测率统计　(%)

	$f_{\mathrm{DOS}}(x)$	$f_{\mathrm{U2R}}(x)$	$f_{\mathrm{R2L}}(x)$	$f_{\mathrm{PRB}}(x)$
TD_{DOS}	99.48	0.00	0.00	18.15
TD_{U2R}	10.25	35.89	0.00	0.00
TD_{R2L}	0.36	0.00	0.00	0.57
TD_{PRB}	83.171	0.00	0.00	74.13
TD_{NOMAL}	98.17	99.98	100	99.81

表 2　4 个子集的样本个数

1/10 子集	攻击类型类型	连接记录个数
负类样本	DoS	30 390
	U2R	17
	R2L	1
	probe	310
正类样本	normal	20 875

表 3　各个 SVM 的正、负支持向量个数

	支持向量中正样本数	支持向量中负样本数
$f_{\mathrm{DOS}}(x)$	202	190
$f_{\mathrm{U2R}}(x)$	23	17
$f_{\mathrm{R2L}}(x)$	4	1
$f_{\mathrm{PRB}}(x)$	102	98

样本不均衡使得各个 SVM 分类检测器的性能是不同的。表 1 正好反映了各个 SVM 分类检测器的性能，由此可以构造矩阵 $\boldsymbol{Q}_{5\times4}$ ：

$$\boldsymbol{Q}_{5\cdot4} = \begin{bmatrix} q_{11} & q_{12} & q_{13} & q_{14} \\ q_{21} & q_{22} & q_{22} & q_{23} \\ q_{31} & q_{32} & q_{33} & q_{34} \\ q_{31} & q_{42} & q_{43} & q_{44} \\ q_{51} & q_{52} & q_{53} & q_{54} \end{bmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \end{pmatrix} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5)^{\mathrm{T}}$$

矩阵 $\boldsymbol{Q}_{5\times4}$ 中的每个元素 q_{ij} 正好是表中各个单元格的值，表示第i个子集被第j个SVM分类器正确分类的检测率。根据风险尽可能小的原则和类似Bayesian方法[13]的处理，具体的融合规则算法设计如下：

(1)由 $\boldsymbol{U} = [u_1, u_2, u_3, u_4]$ 得 $\boldsymbol{V} = [1-u_1, 1-u_2, 1-u_3, 1-u_4] = [v_1, v_2, v_3, v_4]$，其中，$\boldsymbol{U}$ 代表了判断为正常类的向量；$\boldsymbol{V}$ 代表了判断为异常类的向量；

(2)由 $\boldsymbol{V} = [v_1, v_2, v_3, v_4]$ 和 $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ 分别计算 $m^i = \boldsymbol{U} \cdot I_{\alpha_i} = [m_1^i, m_2^i, m_3^i, m_4^i]$，其中，$i = 1,2,3,4$，$I_{\alpha_i}$ 表示对角线元素为 α_i 的各个分量；其余元素为 0 的 4 阶方阵；

(3)计算 $\beta_i = \max\{m_1^i, m_2^i, m_3^i, m_4^i\}, i = 1,2,3,4$，计算 $\gamma = \max\{\beta_1, \beta_2, \beta_3, \beta_4\}$；

(4)由 $\boldsymbol{U} = [u_1, u_2, u_3, u_4]$ 和 α_5 计算计算 $m^5 = \boldsymbol{U} \cdot I_{\alpha_5} = [m_1^5, m_2^5, m_3^5, m_4^5]$，$\eta = \min\{m_1^5, m_2^5, m_3^5, m_4^5\}$；

(5)判决结果：$f(U) = \begin{cases} 1, & \text{if} \quad \eta - \gamma \geq \delta \\ -1, & \text{if} \quad \eta - \gamma < \delta \end{cases}$，其中，$\delta$ 是初试设置的表示区分度的值，一般设置为接近 0 的非负值，比如 $\delta = 0.1$。

4　实验结果及分析

下面从两个方面的实验结果分析以上模型算法的性能：

(1)用前面提到的 KDD99 标签集中分别划分出来的异常子集 $TD_{\mathrm{DOS}}, TD_{\mathrm{U2R}}, TD_{\mathrm{R2L}}, TD_{\mathrm{PRB}}$ 和正常子集 TD_{NOMAL} 分别作为实验的测试数据集，取算法中 δ =0，这 5 个测试子集的检测率分别为 99.51%, 30.77%, 0.65%, 95.67%, 98.13%。和前面表 2 中任一列的值对比，可以发现，除了对 TD_{NOMAL} 检测率有少量下降外，其余的检测率都有了明显的改善。

(2)直接用 KDD99 的标签集，对其测试统计总共的检测性能，得到检测正确率、虚警率、误报率如表 4 所示(近似认为将+1 类判断成-1 类为虚报，将-1 类判断为+1 类为误报)。

表 4　带有正确分类标签子集的检测率　(%)

	检测率	虚警率	误警率
$\delta = 0$	97.15	2.31	0.36
$\delta = 0.1$	97.15	2.30	0.37
$\delta = 0.2$	97.14	2.30	0.38
$\delta = 0.3$	97.14	2.29	0.38

为和标准的集中式 SVM 检测性能做比较，也用 KDD 99 的 1/10 子集作为训练集，进行一次性训练得到支持向量和判别函数，用 KDD99 的标签集做测试，得到正确检测率为 81.17%、虚警率为 18.57%、误报率为 0.26%，和表 4 中得到的结果相比，采用融合方法的检测精度提高的同时降低了虚警率，代价是误警率有了很少的增加。从危险性考虑，少许的误警率的增加换来明显的检测精度提高和明显的虚警率降低，这是非常值得的。

5　结束语

从以上两个方面实验来看，采用针对局部信息检测然后在决策中心融合的方法，可以很好地改善检测攻击的效果。这种构架的模型还有另外的优点：(1)可以找到具体的攻击检测效率的一些细节信息，从而指出改进的方向，比如，只有一个数据大小负类样本，以此训练得到的 SVM 分类器 $f_{R2L}(x)$ 的性能会很差，因此，补充 TD_{U2R} 中的样本信息，可以改善检测性能。(2)更新方便，由于各个局部 SVM 分类检测器之间互相独立，因此分开训练、分开更新。这都是集中式构架中无法达到的。

参考文献

[1] 王　勇. 基于计算智能的分布式入侵检测方法研究[D]. 上海: 华东理工大学, 2005.

[2] 王帅伟. 支持向量机在入侵检测中的应用研究[D]. 桂林: 桂林电子工业学院, 2005.

[3] 唐正军. 网络入侵检测系统的设计与实现[M]. 1 版. 北京: 电子工业出版社, 2002: 2-5.

[4] 唐　豫. 基于数据挖掘的分布式入侵检测系统的研究与实现[D]. 桂林: 桂林电子科技大学, 2006.

[5] Mukkamala S, Janoski G,, Sung A. Intrusion Detection Using Neural Networks and Support Vector Machines[C]//Proc. of the 2002 International Joint Conference on Neural Networks. [S. l.]: IEEE Press, 2002: 1702-1707.

[6] Kung S Y, Mak M W, Lin S H. Biometric Authentication: A Machine Learning Approach[M]. [S. l.]: Prentice Hall PTR, 2004.

[7] Lin Chih-jen, Ruby C Weng. Simple Probabilistic Predictions for Support Vector Regression[Z]. (2004-09-09). http://www.csie.ntu.edu.tw/~ cjlin/papers/svrprob.pdf .

[8] Wu Ting-fan, Lin Chih-jen, Ruby C Weng. Probability Estimates for Multi-class Classification. (2004-09-09). http://www.csie.ntu.edu.tw/~cjlin/ papers/svmprob/ svmprob.pdf.

[9] KDD Cup 1999 Data[Z]. (1999-08-08). http://kdd.ics.uci.edu/databases/ kddcup99/kddcup99.html.

[10] Zang Xizhe, Zhao Jie, Wang Cher. Study on a SVM-based Data Fusion Method[C]//Proc. of 2004 IEEE Conference on Robotics, Automation and Mechatronics. Singapore: [s. n.], 2004: 413-415.

[11] Lie Wennung, Su Chenkang. News Video Classification Based on Multi-modal Information Fusion Image Processing[C]//Proc. of 2005 IEEE International Conference on ICIP. [S. l.]: IEEE Press, 2005 .

[12] 王　勇, 王行愚, 张瑞霞. 基于 D-S 证据理论的分布式入侵检测方法研究[J]. 计算机工程与应用, 2004, 40(13).

[13] 孙即祥. 现代模式识别[M]. 1 版. 长沙: 国防科技大学出版社, 2002: 358-365.

[14] 邓乃扬, 田英杰. 数据挖掘中的新方法——支持向量机[M]. 1 版. 北京: 科学出版社, 2004: 146-149, 152.

(上接第 150 页)

…….

--- 2001:250:f007::1 ping statistics ---

229 packets transmitted, 227 received, 0% packet loss, time 23740ms

rtt min/avg/max/mdev = 2.527/1527.387/7646.329/2340.050 ms, pipe 75

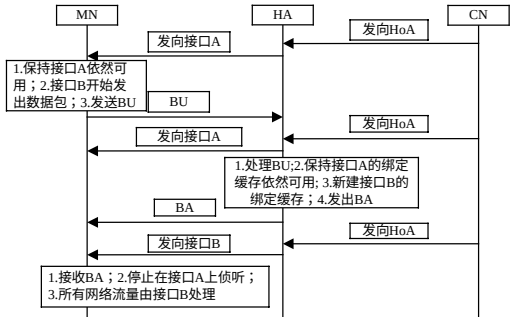

图 3　软切换流程

其中，MN 的家乡地址为 2001: 250: f006: 0:86:100:6256: 5533；CN 地址为 2001:250: f007::1，执行了 2 次切换，从 35 号 ICMPv6 包开始，MN 从 WLAN 切换到 CDMA 中，从 121 号 ICMPv6 包开始，MN 从 CDMA 切换回 WLAN。因为 WLAN 带宽较大，所以当 MN 切换到 CDMA 后，网络中已经没有发向 WLAN 接口的 ICMPv6 包。而当 MN 从 CDMA 切换到 WLAN 时，由于 CDMA 相对较大的延迟，因此当切换完成后，网络上仍然有发向 Silkroad 接口的 ICMPv6 包，这时可以看到这些包能被依然可用的 Silkroad 接口接收，其延迟体现出该包是经过 CDMA 网络传送的。最后整个切换过程没有任何丢包，实现了软切换的预期目标。

4　结束语

IPv6 取代 IPv4 是互联网发展的趋势，而移动 IPv6 作为 IPv6 的一部分，是一种良好的终端移动解决方案，虽然目前在安全性等方面还存在一些缺陷，但是随着研究的深入，整个移动 IPv6 协议机制将得到进一步完善，并很有可能成为未来 IPv6 世界的主要移动技术。本方案使移动终端在目前的网络设施条件下利用移动 IPv6 实现 CDMA 和 WLAN 之间的垂直切换，并针对 CDMA 的具体状况进行了优化，使得 ICP 能提供更多针对移动用户的增值服务，逐步培养用户的认知度和消费习惯，为移动 IPv6 的大规模应用打下广泛的用户基础。

参考文献

[1] 代　刚, 马　严. 移动 IPv6 技术的研究及其在 Linux 环境下的实现[EB/OL]. (2002-03-02). http://scholar.ilib.cn/Abstract.aspx?A=zxtxjs200203007.

[2] Johnson D, Perking C, Arkko J. Mobility Support in IPv6[S]. RFC 3775-2004, 2004-06.

[3] 谢　晨, 吴中福. 移动 IPv6 原理及基于 MIPL 的实现[J]. 计算机科学, 2005, 32(4): 121-124.