第 33 卷 第 1 期 计 算 机 工 程 2007 年 1 月
Vol.33 No.1 Computer Engineering January 2007
·安全技术· 文章编号：1000—3428(2007)01—0153—03 文献标识码：A 中图分类号：TP393.08

基于 SVM 主动学习的入侵检测系统

段丹青[1,2]，陈松乔[1]，杨卫平[1,2]
(1. 中南大学信息科学与工程学院，长沙 410083；2. 湖南公安高等专科学校计算机科学技术系，长沙 410006)

摘 要：研究在入侵检测中，采用基于支持向量机(SVM)的主动学习算法，解决小样本下的机器学习问题。该文提出了基于 SVM 主动学习算法的系统框架及适用于入侵检测系统的 SVM 主动学习算法，讨论了候选样本集的组成比例、候选样本集数量及核函数的不同参数选取对检测结果的影响。通过实验验证，基于 SVM 主动学习算法与传统 SVM 算法相比，能有效地减少学习样本数，提高检测精度。
关键词：入侵检测；支持向量机；主动学习

Intrusion Detection System Based on Support Vector Machine Active Learning

DUAN Danqing[1,2], CHEN Songqiao[1], YANG Weiping[1,2]
(1. College of Information Science and Engineering, Central South University, Changsha 410083;
2. Department of Computer Science and Technology, Hunan Public Security College, Changsha 410006)

【Abstract】 Using support vector machine(SVM) active learning in intrusion detection to resolve the problem of machine learning in the small sample size. This paper provides a framework of intrusion detection system based on SVM active learning, and it also provides a SVM active learning algorithm for intrusion detection system, discusses how the composition of the unlabeled sample set, the size of the unlabeled sample set and the parameter of the kernel function affect the accuracy of the SVM. Compared with the traditional SVM self-learning algorithm, the experiment shows active learning algorithm can immensely reduce the number of the training date and efficiently improve the performance of the classifier in intrusion detection system.
【Key words】 Intrusion detection; Support vector machine(SVM); Active learning

入侵检测技术中的异常检测方法由于能够检测出未知的攻击，因此成为目前入侵检测系统研究的热点。异常检测方法大多采用基于机器学习的方法建模，在学习阶段，需要大量的、完备的训练样本集才能达到较理想的检测性能，同时系统训练时间较长。然而，在现实的网络环境中，完备训练样本集的获取是非常困难的。如何实现在小样本的情况下，获得理想的检测效率成为入侵检测系统研究面临的难题。

为了解决这个问题，本文提出一种基于SVM主动学习算法的网络入侵检测系统。支持向量机(SVM)能较好地解决小样本学习问题，同时具有很好的泛化能力[1]。主动学习算法能根据学习进程，主动选择最佳的样本进行学习，从而能有效减少所需评价样本的数量[2]，大幅缩短训练时间。将SVM主动学习算法应用于网络入侵检测中，可以实现在较少数目学习样本的情况下，保证系统的分类精度不降低甚至有所提高，从而达到在入侵检测系统中提高训练速度和降低构建训练样本集代价的目的，以提高整个系统的入侵检测性能。

1 系统结构

系统结构如图 1 所示，系统由数据采集、特征提取、向量化处理、支持向量库、SVM 训练、SVM 检测及系统响应等模块组成。

数据采集模块捕获网络中的数据流，采用 Tcpdump 实现；特征提取模块从捕获的网络数据包中提取网络连接的特征数据信息；向量化处理模块将这些特征数据转换成 SVM 所需的向量形式，并存入支持向量库。支持向量库中保存了 SVM 训练数据、实时检测数据及检测结果。

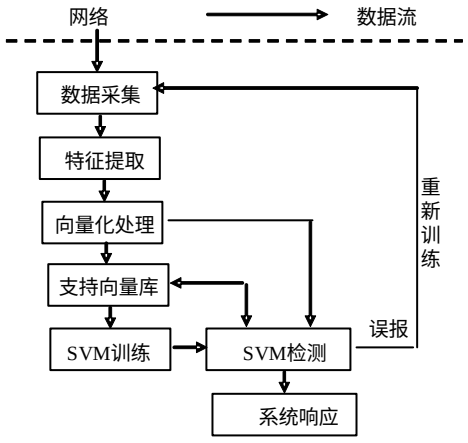

图 1 系统结构

系统的入侵检测引擎采用 SVM 主动学习算法实现，分为两个阶段：第 1 个阶段为训练阶段，使用支持向量库中的训练数据训练 SVM 分类器，训练时采用 SVM 主动学习算法；第 2 阶段为检测阶段，训练过的 SVM 分类器用于对经过数

基金项目：国家自然科学基金资助项目(60403032)；湖南省教育厅青年项目基金资助(03B009)
作者简介：段丹青(1968－)，女，博士生，主研方向：网络安全；陈松乔，教授、博导；杨卫平，硕士生、讲师
收稿日期：2006-03-10 **E-mail：**csddqing@163.com

据预处理的网络数据包进行检测，并将检测结果存入支持向量库。若发现入侵，则调用系统响应模块采取相应的响应策略。若实际检测有误，则进行误差分析，并重新训练 SVM 分类器。SVM 分类器的训练是个不断重复的过程，通过多次训练，使 SVM 分类器的分类精度不断提高。系统详细的检测流程如图 2 所示。

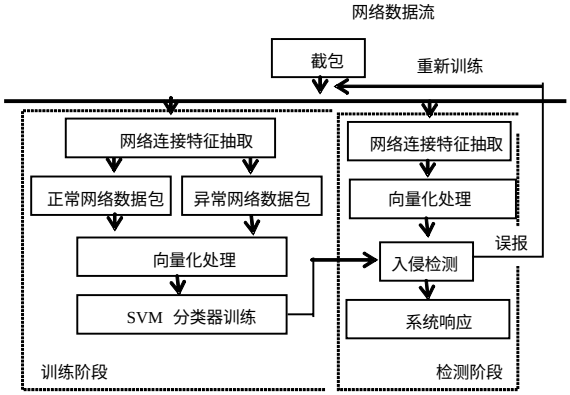

图 2 系统的入侵检测逻辑流程

2 SVM 分类原理

SVM 是由 Vapnik 等人于 1995 年提出的从统计学习理论发展出的一种模式识别方法，是目前较为流行的适用于小样本训练的大边缘分类器。入侵检测可看成是一个二分类问题，通过 SVM 分类器将所有网络数据包分为正常和异常两种。

在输入空间 X 中，把向量 $X=(x_1,x_2,\ldots,x_n)$ 看成一个网络连接集，每一个网络连接 x_i 包含 d 个特征，这些特征是从网络数据包中提取出来的能表示是否入侵的信息。对于每一个网络连接 x_i，相应的输出标注为 y_i，对于正常的网络连接，y_i 标注为 1；对于异常的网络连接，y_i 标注为–1。SVM 用于分类问题就是寻找一个最优分类超平面，它不但可以将给定的输入样本正确地划分为正常和异常两类，而且使得被分成的两类数据间的分类间隔尽可能大。

2.1 线性情况

当训练样本集线性可分时，分类超平面的描述为

$$w\bullet x+b=0 \tag{1}$$

式(1)中向量 w 为分类超平面的权系数，b 是分类阈值。

最优分类超平面可通过解下面的凸二次优化问题获得：

$$\min\phi(w)=\|w\|^2/2 \tag{2}$$

约束条件为

$$y_i(w\bullet x_i)+b-1\geq 0, i=1,2,\ldots,n \tag{3}$$

通过求解，得到最优分类超平面的分类判别函数为

$$f(x)=\mathrm{sgn}[\sum_{i=1}^{n}\alpha_i y_i(x_i\bullet x)+b] \tag{4}$$

其中 α_i 为拉氏乘子，拉氏乘子不为 0 的解向量为支持向量。

2.2 非线性情况

对于非线性问题，SVM$_S$通过选择适当的非线性变换，将输入空间 X 中的训练样本映射到某个高维特征空间 F，使得在目标高维空间中这些样本线性可分。

根据泛函的有关理论，若核函数 $K(x,x_i)$ 满足Mercer条件，它就对应某一变换空间中的内积 $\langle\phi(x_i)\cdot\phi(x)\rangle$ [3]，函数 $\phi: X\rightarrow F$ 是一个从非线性输入空间 X 到高维特征空间 F 的映射，所以求映射 $\phi: X\rightarrow F$ 只要知道如何由输入 x、x_i 计算内积 $\langle\phi(x_i)\cdot\phi(x)\rangle$ 即可，由

$$K(x_i,x)=\phi(x_i)\cdot\phi(x) \tag{5}$$

将式(4)重写，即可得到对应高维空间的分类函数为

$$f(x)=\mathrm{sgn}[\sum_{i=1}^{n}\alpha_i y_i K(x_i\bullet x)+b] \tag{6}$$

这样，在高维空间的内积运算，转化为低维输入空间中的一个简单函数计算。分类函数类型为式(6)的学习机称为支持向量机。

3 SVM 主动学习算法

根据 SVM 分类原理，对于非线性向量的网络连接样本集，可以通过选择适当的核函数，将低维空间中的训练样本映射到某个高维特征空间 F，使得在目标高维空间中这些样本线性可分，从而将样本集划分为正常和异常两类。

但是，目前 SVM 训练算法速度都比较慢，这是因为训练样本的数量决定了二次规划问题目标函数中矩阵的维数，使得求解规划问题的速度与维数呈指数增长。为了提高训练速度，减少学习样本数，缩短训练时间，将主动学习引入 SVM 中。主动学习与传统的被动学习的本质区别在于，它可以在候选的样本集中，主动选择对于当前分类器不确定度最大的新样本进行训练，来进一步设计新的分类器，从而使得可以用尽可能少的标注样本数来实现尽可能高的分类精度。

SVM 主动学习机包括两个独立的部分 (f,q)，f 是一个 SVM 分类器，$f: X\rightarrow\{-1,1\}$ 使用训练样本集进行学习，q 是一个查询函数，根据训练样本集，决定下一步应从候选集 U 中选择哪一个样本进行标注。

SVM 主动学习机由查询函数 q 采取某种查询策略，从未标注的候选样本集 U 中选择下一个应标注的样本。根据泛函原理可知，对于线性可分问题，分类间隔中的样本对分类器的影响较大。因此，本文中 q 采用的查询策略为：每次选择离分类面最近的一个样本作为新样本进行训练。采用这种策略，每次选择进行学习的样本都是不确定性最大的样本，它对分类器的影响也最大，候选样本集中剩下的样本对分类器的影响逐步减弱。这种策略充分体现了 SVM 的本质，即分类器仅与支持向量有关，与其它向量无关。

下面给出本文采用的 SVM 主动学习算法：

输入：未带类别标注的候选样本集 U，每次从 U 中采样个数为 1

输出：分类器 f，预标注样本

(1)从候选样本集 U 中选择 i 个样本并正确标注其类别，构造初始训练样本集 T，使 T 中至少包含一个输出 y 为 1 和一个输出 y 为–1 的样本；

(2)根据训练集 T 构造 SVM 分类器 f；

(3)对 U 中所有样本使用 f，标注为(x, $\hat{y}$)，其中 $\hat{y}$ 为分类器 f 给向量 x 预先打上的标注；

(4)从样本集 U 中选择一个离分类边界最近的未标注样本(x, $\hat{y}$)；

(5)将该样本正确标注后加入训练集 T 中(y 为 x 的正确标注)；

(6)若检测精度达到某一设定值，算法终止，返回 f；否则重复第(2)步。

4 数据预处理

对于系统采集到的网络数据流，提取相关特征进行检测。在特征提取时主要考虑与网络连接相关的特征，这些特征最能体现出该连接的类型。对于每一次网络连接，抽取 18 个特

征，分别是：duration(连接持续的时间)，protocol_type(协议类型)，service(目标网络服务)，src_bytes(源地址到目标地址的字节数)，dst_bytes(目标地址到源地址的字节数)，flag(网络连接状态，正常或错误)等。

由特征提取模块获得的网络连接记录信息格式复杂，必须将这些信息转换成 SVM 能够处理的向量形式。对于数据的向量化处理，遵循文献[4]提出的规则。第 1 步，先将所有类型的数据转换成以二进制表示的数字形式。转换时采用文献[5]提出的基于距离度量函数 HVDM 的方法，对数据进行归一化处理。例如：对于协议类型的特征值有{tcp,udp,icmp}，则转换成二进制分别为{1,0,0}、{0,1,0}、{0,0,1}。第 2 步，对这些特征值的范围进行处理，使得每类特征数据的取值范围在区间[0,1]中。这样处理一方面可以避免取值范围大的特征支配那些取值范围小的特征；另一方面可以降低机器的计算时间。

5 实验结果及分析

5.1 数据源

实验中采用的数据取自 1999 年 DARPA 为 KDD(知识发现与数据挖掘)竞赛提供的一个异常检测的标准数据集，该数据集包括约 500 万条训练集和 300 万条测试集，数据中包括 4 种类型的攻击：DoS(拒绝服务攻击)，R2L(未经授权的远程访问)，U2R(对本地超级用户的非法访问)和 Probing(扫描与探测)。

5.2 实验结果及分析

通过测试比较，实验选用的核函数为高斯核函数 $k(x_i, x) = e^{-\gamma \|x_i - x\|^2}$，采用检测精度、误报率和漏报率作为衡量入侵检测系统性能的 3 个重要指标，主要研究不同候选样本组成、样本数量、参数选择对检测效果的影响，从而找出系统性能最佳的工作点。

实验 1 候选样本组成比例的选择

本实验主要研究不同组成比例的候选样本集对检测结果的影响。从 KDD 的训练样本集中选取 200 个训练数据作为候选样本，改变候选样本集的组成，如表 1 所示；同时从测试集中选取了 1 000 个样本作为测试集，其样本组成与候选样本集相同，以保证数据的独立同分布性。实验中核函数控制因子 γ 值设定为 1，错分惩罚因子 C 设为 100。实验结果如表 2 所示。

表 1 候选样本集组成

U 中样本总数	正常样本与异常样本比
200	90%：10%
200	70%：30%
200	50%：50%
200	30%：70%

表 2 实验 1 结果

正常样本与异常样本比	检测精度(%)	误报率(%)	漏报率(%)
90%：10%	95.21	0.87	1.35
70%：30%	96.89	0.43	0.98
50%：50%	97.42	0.22	0.63
30%：70%	95.79	0.39	1.17

实验表明，当候选样本集中正常样本与异常样本比为 1：1 时，检测效果最好。这是因为候选样本集中的数据偏斜较大时，对于数量较少的数据，学习机缺乏足够的学习样本来正确识别。当候选样本集中的正负样本数相当时，SVM 学习机分类精度较高。

实验 2 候选样本数量的选择

本实验主要研究候选样本数量对检测结果的影响。根据实验 1 结论，候选样本集中正负样本比例固定为 1：1，改变候选样本集的规模，如表 3 所示；测试集中样本数为 1 000，样本组成与候选样本集相同，γ =1，C=100。表 4 为实验 2 结果。

表 3 候选样本集组成

U 中样本总数	正常样本与异常样本比
200	50%：50%
500	50%：50%
1 000	50%：50%

表 4 实验 2 结果

U 中样本总数	检测精度(%)	误报率(%)	漏报率(%)
200	97.42	0.87	1.35
500	99.56	0.63	0.92
1 000	99.85	0.51	0.88

从表 4 可以看出，当候选样本集中样本数越多，分类器的检测精度越高，这是因为样本数量越多，代表正常和异常之间的区别的数据越多，从而使得 SVM 的检测精度随着学习样本的增多而不断提高。当样本数从 200 增加到 500 时，检测精度提高幅度较大；而样本数从 500 增加到 1 000 时，系统检测精度的提高不明显，但系统的存储空间及训练时间却随着样本数的增加成指数级增长，因而当候选样本集为 500 时，已完全能够满足系统检测精度的需要。

实验 3 参数的选择

本实验主要研究核函数的不同参数取值对检测结果的影响。根据实验 1、实验 2 的结论，选取 500 个训练数据作为候选样本，候选样本集中正负样本比例为 1：1；测试集中样本数为 1 000，样本组成与候选样本集相同。改变参数 γ 和 C 的取值，测试在不同组合下 SVM 的分类性能。实验结果如表 5 所示。

表 5 实验 3 结果

C	γ	检测精度(%)	误报率(%)	漏报率(%)
50	0.5	98.32	0.64	0.57
50	1	97.56	0.76	0.96
50	2	97.13	0.92	1.26
100	0.5	99.32	0.37	0.45
100	1	99.03	0.63	0.79
100	2	98.17	0.81	0.95
200	0.5	97.42	0.89	0.67
200	1	96.71	0.92	0.84
200	2	96.32	1.02	1.43

实验 3 结果表明，当 C=100，γ =0.5 时，检测精度最高。

5.3 SVM 主动学习算法与 SVM 传统算法检测精度的比较

为了验证 SVM 主动学习算法可以较大幅度地减少学习样本数，有效地提高系统速度，将其与 SVM 传统算法进行了对比实验。

在本实验中采用 5.2 节的结果，候选样本集为 500，正负样本比为 1：1，检测样本集为 1 000，样本组成与候选样本相同。核函数的参数取值为 C=100，γ =0.5。实验结果如图 3 所示。

从图 3 中可以看出，与 SVM 传统算法相比，SVM 主动学习算法的检测精度比 SVM 传统学习算法高；当标注样本数约为 80 时，SVM 主动学习算法的检测率已达到 98.9%，

（下转第 180 页）