文章编号:1001-506X(2012)05-1036-05

基于修正核函数 SVM 的网络入侵检测

井小沛,汪厚祥,聂 凯
(海军工程大学电子工程学院,湖北 武汉 430033)

摘 要:支持向量机分类方法在小样本、非线性情况下具有较好的泛化性能,在入侵检测系统中有着广泛的应用。针对入侵检测过程中可能出现的由两类样本不平衡造成的分离超平面偏移现象,以核函数所蕴含的黎曼几何为依据,引入一个伪一致性变换函数,对核函数进行修改,提高支持向量机的分类泛化能力,建立基于支持向量机的网络入侵检测系统,并对系统总体结构和运行机制进行了详细的描述。实验仿真表明,该系统可有效地提高入侵检测的准确率,改善由于数据集不平衡造成的支持向量机分类偏移的情况。

关键词:入侵检测;支持向量机;修正核函数;不平衡数据;黎曼几何
中图分类号:TP 393 **文献标志码**:A **DOI**:10.3969/j.issn.1001-506X.2012.05.32

Network intrusion detection based on modified kernel function SVM

JING Xiao-pei, WANG Hou-xiang, NIE Kai
(*Electronic Engineering Institute, Naval University of Engineering, Wuhan 430033, China*)

Abstract: As the support vector machine (SVM) classification approach has a good generalization performance in the cases of small number and non-linear samples, it is widely used in network intrusion detection fields. In order to resolve the offset phenomenon of separating a hyperplane caused by imbalanced data, Riemannian geometry inherent in a nuclear function is regarded as an important basis and a pseudo-consistency transformation function is also introduced, both of which are used to modify the kernel function and improve the generalization ability of SVM classification. On this basis, an intrusion detection system based on modified kernel function SVM is established, and a detailed description of the overall structure of the system and operating mechanism is made. Finally, simulation experiment shows that this system can achieve a more accurate detection rate and improve the SVM's classification offset phenomenon caused by imbalanced data sets.

Keywords: intrusion detection; support vector machine (SVM); modified kernel function; imbalanced data; Riemannian geometry

0 引 言

入侵检测技术是一种主动保护自己的网络和系统免遭非法攻击的网络安全技术。它从计算机系统或者网络中收集、分析信息,检测任何企图破坏计算机资源的完整性、保密性和可用性的行为,并做出相应的反应。为了提高入侵检测的准确率,入侵检测领域的研究者引入各种新的方法,如人工免疫[1]、遗传算法[2]、神经网络[3]等,力图建立更为精确的检测模型,来提高入侵检测的检测率,降低误报率。但是以上这些方法都有一个共同点,就是学习算法多是基于全样本数目的,要求学习样本数目足够多足够全。而在实际应用中,由于多方面的原因,这一要求往往得不到保证。

支持向量机(support vector machine, SVM)是统计学习理论的产物,针对有限样本情况,SVM 建立了一套完整的、规范的基于统计的机器学习理论和方法,而且对数据的维数和多变性不敏感,具有较好的分类精度和泛化能力。因此,将 SVM 应用到入侵过程中,可以在样本数目有限和线性不可分的情况下取得较好的分类效果,达到较高的检测率。很多研究人员把 SVM 应用到入侵检测系统中去:文献[4]将特征提取和 SVM 应用到入侵检测中,首先对网络数据的特征进行分析,提取关键特征,在此基础上建立 SVM 分类模型,从而达到提高检测效率和减少检测时间的目的;文献[5]将最近邻和 SVM 应用到入侵检测系统中,通过计算样本中同异类点的距离,对样本进行修剪,提高 SVM 学习算法的效率;文献[6]从几何角度来解释 SVM,

收稿日期:2011-07-26; **修回日期**:2012-03-15。
基金项目:海军十一五预研项目基金(4010601010201)资助课题
作者简介:井小沛(1983-),男,博士研究生,主要研究方向为信息安全。E-mail:jingxiaopei@163.com

将并行凸包分解计算用来提取训练样本空间几何凸包定点,从而达到约简 SVM 训练样本的目的;文献[7]则针对网络数据的不断变化,将增量 SVM 应用到入侵检测中,并且针对样本数据变化带来的分类超平面振荡现象,通过建立支持向量储备集,将可能成为支持向量的样本数据放到该数据集中,以降低样本变化带来的分类超平面振荡问题。以上研究都是基于入侵检测训练集中正常样本和异常样本基本相等的情况,没有考虑到样本数据集不平衡的情况。本文在研究上述文献的基础上,考虑到入侵检测系统中正常样本和异常样本都是实时变化,且两类样本数目可能相差悬殊等情况,将基于不平衡数据的 SVM 应用于入侵检测领域,从几何角度出发,将一种基于修正核函数的 SVM 模型应用到入侵检测系统中,使系统更好地适应入侵检测实时性和自适应性的要求。

1　分类支持向量机

对于给定分类问题,设定其训练样本集为 $\{x_i, y_i\}(i=1,2,\cdots,l)$, $x_i \in X$, $y_i \in \{-1,+1\}$,其中 X 为 d_n 维空间。SVM 即为线性分类器,通过构造最优分类面,使得类别间的分类间隔最大。但是这样的分类面不是唯一的,SVM 训练算法就是求出与两类样本间隔尽可能大的分类面。因此,将求最优分类面的问题转化为优化问题,表示为

$$\begin{cases} \min\ \phi(\omega) = \dfrac{1}{2}\|\omega\|^2 + C\sum\limits_{i=1}^{l}\xi_i \\ \text{s.t.}\ \ y_i(\omega K(x\cdot x_i)+b) \geqslant 1-\xi_i \end{cases} \tag{1}$$

式中,ξ_i 为松弛变量,$\xi_i \geqslant 0$;C 为惩罚系数;$K(x\cdot x_i)$ 为核函数,它的作用是将低维非线性空间的数据通过核函数映射到高维属性空间(也称特征空间),将非线性变换转换为某个高维空间中的线性问题[8]。

其对偶问题为

$$\begin{cases} \min\ Q(\alpha) = \dfrac{1}{2}\sum\limits_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j K(x_i\cdot x_j) - \sum\limits_{i=1}^{l}\alpha_i \\ \text{s.t.}\ \ 0 \leqslant \alpha_i \leqslant C,\ i=1,2,\cdots,l \\ \qquad \sum\limits_{i=1}^{l}\alpha_i y_i = 0 \end{cases} \tag{2}$$

2　基于不平衡数据的修正核函数 SVM 模型

2.1　入侵检测数据集的不平衡现象

在入侵检测过程中,正常行为模式的数量和异常行为模式的数量都是动态变化的,不可能保证两类样本模式是严格相等的。尤其是在出现以下两种情况时:①网络正常运行过程中,系统正常行为模式占绝对多数;②网络遭受攻击中,系统异常行为模式占绝对多数。

当两类样本数目不平衡时,就会使 SVM 分类超平面明显偏向样本少的一方,造成较大的分类误差[9]。因此,为了提高检测效率,必须要解决数据集不平衡带来的分类超平面偏移现象。目前主要有两种方法来解决样本数据集不平衡带来的分类超平面偏移问题:一是对样本多的一方进行欠采样,或是对样本少的一方进行过采样。文献[10]采用对少数样本进行过采样的方法提高分类精度,而文献[11]将过采样和欠采样两种方法结合在一起。二是对分类方法进行修改,来适应样本的变化。例如文献[12]对两类样本采用不同的惩罚因子,文献[13]则通过计算两类样本的数目差来修改核函数,达到扩大分类间隔的目的。

本文针对实验采用的 KDD CUP 99 数据集的特点:样本数目的多少直接决定样本分布空间的密度,而对空间的大小影响不大,也就是说样本数目越多,样本的区域越稠密(验证实验结果见表 1)。采用对核函数进行修正的方法来解决样本不平衡带来的分类面偏移现象,通过对核函数所蕴含的几何度量的深入分析,导出了本文采用的高斯径向基函数(radical basis function, RBF)的核函数($K(x,x')=e^{-\|x-x'\|^2/(2\sigma^2)}$)的黎曼度量,并引入修正函数来改进黎曼度量,以达到提高分类效果的目的。

表 1　不同数目样本集的聚类半径

聚类半径	数据集样本数				
	1 000	2 000	3 000	4 000	5 000
正常样本	2.540 2	2.533 5	2.611 2	2.613 5	2.613 1
异常样本	2.677 3	2.655 2	2.650 7	2.677 5	2.678 6

从表 1 可以看出随着样本数目的增多,不管是正常样本还是异常样本,聚类半径基本不变。

2.2　核函数的空间特性

支持向量方法中,核函数将数据从 d_n 维空间 X 映射到 d_F 维的特征空间 H 中。也就是满足 Mercer 条件的对称、连续核,存在 Hilbert 空间 H,映射 Φ: $X \to H$,和展开式

$$K(x,x') = \sum_{n}^{d_F}\Phi_n(x)\Phi_n(x') \tag{3}$$

定义 L_S 表示数据存在的空间,它是 d_n 维空间的子集,并假设数据是连续的。让 S 表示在核函数作用下 L_S 的像。那么在特征空间 H 中,S 是一个可微流形[13]。对于 L_S 上的两点 x_i 和 x_j,它们在面 S 上的距离即黎曼距离定义为

$$ds^2 = \sum_{i,j} g_{ij}(x)dx_i dx_j \tag{4}$$

式中,$g_{ij}(x)$ 为空间 H 上的黎曼度量。于是特征空间 H 成为黎曼空间,其体积为

$$dv = \sqrt{g(x)}dx_1\cdots dx_n \tag{5}$$

式中,$g(x)=\det(g_{ij}(x))$。直观地说,$g(x)$ 反映了特征空间中点 $\Phi(x)$ 附近局部区域被缩放的程度。因此,也称 $g(x)$ 为缩放因子。由式(3)可以得出

$$g_{ij}(x) = \frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j'}K(x,x') \tag{6}$$

对高斯 RBF,有 $g_{ij}(x)=\delta_{ij}/\sigma^2$,这里 δ_{ij} 是 Kronecker δ[14]。

因此,从高斯核的核映射和黎曼特性,可以找到提高高斯核性能的方法:为有效地将两类不同模式区分开,应尽量拉大它们之间距离。

2.3 修正的高斯 RBF 核函数

由 2.1 节的分析知道，对于入侵检测数据集，当样本数目不等时，存在样本稠密区和样本稀疏区。而采用全局核宽度的高斯核诱导的度量在样本的稠密区域，会存在过学习现象；在样本的稀疏区域，会存在欠学习现象。因此，对不同样本采用不同黎曼度量的方法，来实现 SVM 的全局优化策略和局部适应性之间折中就显得很有必要。本文采用修正核函数的办法解决这个问题。这里以定义的形式给出修正核函数的组成[15]。

定义 1 对于一个正的可微标量函数 $C(x)$，定义

$$\tilde{K}(x,x') = C(x)C(x')K(x,x') \tag{7}$$

称为核函数通过因子 $C(x)$ 的保角变换，则 $\tilde{K}(x,x')$ 成为 SVM 的修正核函数。

根据定义 1，非线性映射 Φ 被修正为 $\tilde{\Phi}(x)=C(x)\Phi(x)$。这时黎曼度量 $g_{ij}(x)$ 变为

$$\tilde{g}_{ij}(x) = C^2(x)g_{ij}(x) + C_i(x)C_j(x) \tag{8}$$

式中，$C_i(x)=\partial C(x)/\partial x_i$。

修正函数的选取应考虑以下两个因素：①在样本的稠密区域，减小核数值，在稀疏区域，增加核数值，从而达到间接影响黎曼度量的目的；②对类边界样本进行体积扩张，即尽量放大黎曼度量。

设 d_n 维空间 X 中存在两类样本集 S_+ 和 S_-，样本数目分别是 N_1 和 N_2。构造修正函数之前先用 $d(x)$ 表示一个样本与同类其它样本的欧式距离平均值，可以断定在稠密区域样本的 $d(x)$ 值会小于在稀疏区域样本的 $d(x)$ 值。因此，$d(x)$ 是空间分布变化的一个近似量化指示。此外，那些距离样本类中心较远的样本的 $d(x)$ 也比距离样本类中心较近样本的大。$d(x)$ 通过下式计算

$$d(x) = \begin{cases} \left(\sum_{i=1}^{N_+} \sqrt{\| x - x_i^+ \|^2}\right)/N_1, & x \in S_+ \\ \left(\sum_{i=1}^{N_-} \sqrt{\| x - x_i^- \|^2}\right)/N_2, & x \in S_- \end{cases} \tag{9}$$

根据式(9)，构造修正函数为

$$C(x) = \exp\left(d(x)/r_i\right) \tag{10}$$

式中，r_i 为样本类的聚类半径。

将式(10)代入式(7)就得到修正核函数。由此可见，采用本文构建的修正核函数，稀疏样本和边界样本都能获得较大的体积扩张，有利于提高分类精度。

本文的修正核函数算法具体步骤如下：

(1) 通过 K 均值聚类算法对样本进行聚类分析，计算出两类的聚类中心，由此求出聚类半径 r_i；

(2) 根据式(9)和式(10)计算修正函数；

(3) 将修正函数代入式(7)得到修正核函数。

建立修正核函数后，就可以对训练集进行训练建立 SVM 模型。

3 基于 SVM 的入侵检测系统

本文设计的基于 SVM 的入侵检测系统，通过 SVM 分类技术，把待测数据识别成正常和异常两类。应用 SVM 的入侵检测模型的总体结构如图 1 所示。应用 SVM 进行入侵检测分为两个阶段：模型训练阶段和入侵检测阶段。

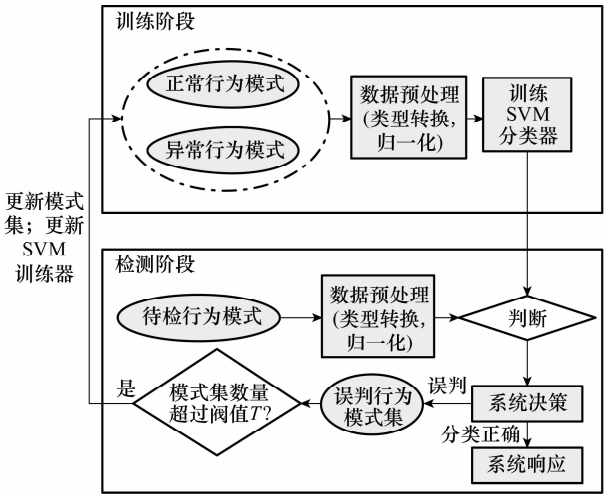

图 1 应用 SVM 的入侵检测模型的总体结构

在模型训练阶段，又可以分为两个阶段：初始训练阶段和更新训练阶段。在初始训练阶段，由先验信息得到训练样本，包括正常行为模式样本和异常行为模式样本。采用参数寻优法找出最优参数使得 SVM 对于训练样本的分类性能达到最优。在更新训练阶段，系统会根据训练样本集的变化，重新对 SVM 模型进行训练，更新模型参数，以适应新的检测环境，提高检测效率。

入侵检测阶段由数据预处理、SVM 分类和判决模块组成。对待检数据进行预处理，然后由训练阶段产生的 SVM 分类器进行分类。并将分类结果送到系统决策模块，如果判断有误，则将该行为模式放进误判行为模式集中。

4 实验仿真分析

4.1 数据预处理

实验数据选用了 1999 年 DARPA 为 KDD 竞赛提供的 KDD CUP 99 数据集。这个数据集是网络入侵检测的标准测试集，为入侵检测研究人员提供训练和测试数据集，以期比较不同入侵检测方法的优劣。数据中包含 41 维特征，包含 34 个数值型字段和 7 个符号型字段[16]。由于数据的类型比较复杂，有离散类型(比如协议类型 tcp，upd，…)和连续类型(持续时间，发送包数量等)，要分开处理。预处理过程分两个步骤。

①映射。将符号型的数据映射到数值型数据。例如，将 protocol_type 的 tcp，udp，icmp 分别映射成数值 1，2，3。

②归一化。采用归一化函数

$$f(x) = \begin{cases} (x - x_{\min})/(x_{\max} - x_{\min}), & x_{\max} \neq x_{\min} \\ 0, & x_{\max} = x_{\min} \end{cases} \tag{11}$$

这样，将所有数据都调整到[0.0，1.0]区间内。

4.2 仿真实验

仿真实验主要验证本文采用的基于修正核函数 SVM 的入侵检测系统的检测效果。实验采用检测率(true positive rate, TPR)来判断模型分类效果。

定义 2 TPR=正确分类的测试样本个数/总的测试样本个数。

实验所需训练集分为两部分:分别模拟正常状态下和入侵状态下 SVM 模型训练的情况。正常状态下,训练集中异常样本数保持为 1 000 不变,入侵状态下,训练集中正常样本数保持为 1 000 不变,具体组成情况见表 2。测试集设为两个,样本数分别为 3 000 和 5 000,其中正常样本数和异常样本数基本相等。

表 2 实验训练集的组成

实验训练集样本数		训练集 1	训练集 2	训练集 3	训练集 4	训练集 5
正常状态	正常样本	1 000	2 000	3 000	4 000	5 000
	异常样本	1 000	1 000	1 000	1 000	1 000
入侵状态	正常样本	1 000	1 000	1 000	1 000	1 000
	异常样本	1 000	2 000	3 000	4 000	5 000

为了使实验更具针对性,实验设定 3 种情况:一是对训练集应用相同参数 C 值进行训练建模,并利用模型对测试集进行测试分类;二是采用文献[12]的对正常样本和异常样本应用不同惩罚因子 C 值($C_i=C_+,C_-$)进行训练建模(训练集 1 除外),并对测试集进行测试分类;三是采用本文提出的修正核函数方法对训练样本进行训练。

正常状态下 3 种实验的 TPR 对比如表 3 所示。由表 3 中的数据可以看到,实验二和实验三的 TPR 相对于实验一都有较明显的提高,这是因为对不平衡样本采取不同处理机制后,有效地改善了分类面的偏移。

表 3 正常状态下 3 种情况的 TPR %

项目		实验训练集 训练集 1	训练集 2	训练集 3	训练集 4	训练集 5
测试集 1	一	97.53	97.8	97.9	97.87	97.97
	二	—	98.4	98.27	98.17	98.47
	三	99.47	99.4	99.5	99.53	99.53
测试集 2	一	98.3	98.46	98.5	98.44	98.48
	二	—	99.02	98.94	98.88	99.06
	三	99.68	99.7	99.66	99.72	99.7

实验结果还表明,实验三的检测率明显高于实验一和实验二,说明本文采用的修正核函数更好地实现了在空间层面对样本黎曼度量的改进,有效地扩大了两类的分类间隔;且很好地扩大了对稀疏样本黎曼度量,从而抓取更多的近邻先验。与实验二的迭代计算不同惩罚因子的方法相比,在几何空间层面的对核函数进行修正的方法更适合于处理入侵检测数据集。

为了直观地显示修正核函数计算方法对分类面偏移改善的效果,本文计算两类错分样本的比值(error classification rate, ECR)。

定义 3 ECR1=异常类错分样本数/正常类错分样本数。

显然,ECR1 的值越大表明分类面向异常类偏移越严重,ECR1 的值趋近于 1 表明分类面比较合理。正常状态下 3 个实验的 ECR1 值如图 2 所示。

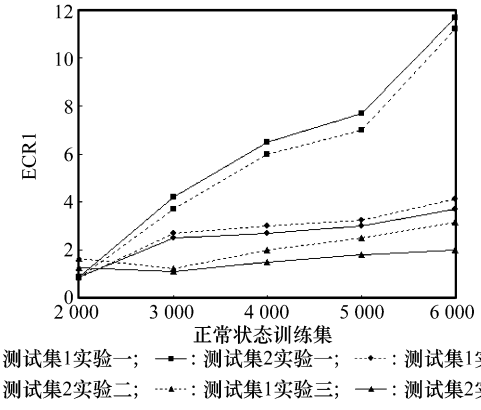

图 2 基于不同惩罚机制的入侵检测模型在测试集上的 ECR1 值

图 2 表明,随着训练集中两类样本数目差距的增大,ECR1 的值就随着增大,表明分类面偏移就越严重。尤其在对两类设定相同惩罚因子的情况下,随着两类样本数目相差悬殊的增大,分类面偏移的变化非常明显,但引入对两类实施不同的惩罚机制后,ECR1 的值变化就比较缓慢,基本上处于稳定状态。可以看到实验三的结果比实验二的结果更理想,说明修正核函数的计算方法更好地解决了分类面偏移的问题。

在入侵状态下,进行与正常状态下相同的 3 种实验。入侵状态下 TPR 对比如表 4 所示。

表 4 入侵状态下 3 种情况的 TPR %

入侵状态 TPR		实验训练集 训练集 1	训练集 2	训练集 3	训练集 4	训练集 5
测试集 1	一	97.53	97.67	97.43	97.4	97.33
	二	—	97.9	97.73	97.87	97.7
	三	99.47	99.37	99.4	99.37	99.33
测试集 2	一	98.3	98.4	98.24	98.2	98.14
	二	—	98.64	98.46	98.56	98.58
	三	99.68	99.58	99.54	99.62	99.65

表 4 中的数据表明对不同样本应用不同的惩罚参数后,模型的检测准确率有较明显的提升。但与表 3 中的数据对比发现,正常状态下的 TPR 高于入侵状态下的 TPR。这是因为正常样本相对比较稳定,有限的正常样本就能代表正常样本的属性特征。而由于入侵行为的多样性,异常样本特征比较多,有限的异常样本不能很好地代表异常样本的属性特征。因此,在入侵检测过程中,尽力收集比较全的入侵特征对模型的入侵检测效率会有较大的提升。

同样,为了直观地显示修正核函数计算方法对入侵状态下分类面偏移改善的效果,定义另一个 ECR。

定义 4 ECR2＝正常类错分样本数/异常类错分样本数。

同样，ECR2 的值趋近于 1 表明分类面比较合理。入侵状态下 3 个实验的 ECR2 值如图 3 所示。

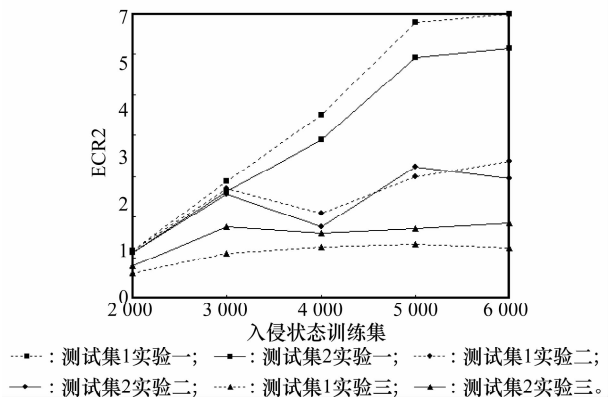

图 3 基于不同惩罚机制的入侵检测模型在测试集上的 ECR2 值

图 3 同样表明，引入对两类实施不同的惩罚机制后，ECR2 的值变化就比较缓慢，基本上处于稳定状态。但与图 2 相比发现，在训练集样本数目相差悬殊时，入侵状态下分类面的偏移没有正常状态下分类面的偏移明显，由此可见，相同数量的异常样本对 SVM 模型分类面的影响没有正常样本明显。

5 结 论

将在小样本、非线性情况下具有较好的泛化性能的 SVM 分类方法应用到入侵检测中，可以保证在先验知识不足的情况下，入侵检测系统仍有较好的分类正确率。但入侵行为是不断变化的，这就要求 SVM 分类模型也必须及时更新来适应新的外部环境。本文详细介绍了系统训练集和模型的更新步骤，特别针对入侵检测过程中训练集中出现的数据不平衡现象，以黎曼几何为理论基础，运用伪一致性变换增大稀疏样本和类边界样本附近的空间体积，使得 SVM 分类间隔增大，来解决由此造成的分类面偏移，提高分类精度。最后通过实验证明，本文构建的 SVM 分类模型在解决样本数据集不平衡方面具有很好的效果，且对行为特征经常变化和攻击行为层出不穷的实时入侵检测系统非常适用。

参考文献：

[1] Chen J Y, Yang D Y, Matsumoto N. A study of detector generation algorithms based on artificial immune in intrusion detection system[J]. *WSEAS Trans. on Biology and Biomedicine*, 2007, 3(4): 29－35.

[2] Yu Y, Huang H. An ensemble approach to intrusion detection based on improved multi-objective genetic algorithm[J]. *Journal of Software*, 2007, 18(6): 1369－1378.

[3] Iren L F, Francisco M P, Francisco J G, et al. Intrusion detection method using neural networks based on the reduction of characteristics[C]// *Proc. of the 10th International Work-Conference on Aartifical Neural Networks*, 2009: 1296－1303.

[4] Xie L X, Zhu D, Yang H Y. Research on SVM based network intrusion detection classification[C]// *Proc. of the 6th International Conference on Fuzzy Systems and Knowledge Discovery*, 2009: 362－366.

[5] 于秋玲. 基于改进 NN-SVM 算法的网络入侵检测[J]. 系统工程理论实践, 2010, 20(1): 126－130. (Yu Q L. Internet intrusion detection system based on improved NN-SVM[J]. *Systems Engineering－Theory & Practice*, 2010, 30(1): 126－130.)

[6] 张雪芹，顾春华. 一种基于约简支持向量机的快速入侵检测模型[J]. 华南理工大学学报(自然科学版), 2011, 39(2): 108－112. (Zhang X Q, Gu C H. A Reduced SVM-based fast intrusion detection model[J]. *Journal of South China University of Technology (Natural Science Edition)*, 2011, 39(2): 108－112.)

[7] Yi Y, Wu J S, Xu W. Incremental SVM based on reserved set for network intrusion detection[J]. *Expert Systems with Applications*, 2011, 38(6): 7698－7707.

[8] Latifur K, Mamoun A, Bhavani T. A new intrusion detection system using support vector machines and hierarchical clustering [J]. *The International Journal on Very Large Data Bases*, 2007, 16(4): 507－521.

[9] Zhao Z Y, Zhong P, Zhao Y H. Learning SVM with weighted maximum margin criterion for classification of imbalanced data [J]. *Mathematical and Computer Modelling*, 2011, 54(3－4): 1093－1099.

[10] Han H, Wang W, Mao B. Borderline-smote: a new over-sampling method in imbalanced data sets learning[C]// *Proc. of the International Conference on Intelligent Computing*, 2005: 878－887.

[11] Liu Y, Yu X H, Huang X J. Combining integrated sampling with SVM ensembles for learning from imbalanced datasets[J]. *Information Processing and Management*, 2011, 47(4): 617－631.

[12] Sun Y, Kamela M, Wongb A. Cost-sensitive boosting for classification of imbalanced data[J]. *Pattern Recognition*, 2007, 40 (12): 3358 － 3378.

[13] Wang B X, Nathalie J. Boosting support vector machines for imbalanced data sets[J]. *Knowledge and Information Systems*, 2010, 25(1): 1－20.

[14] Yendrapalli K, Mukkamala S, Sung A H. Biased support vector machines and kernel methods for intrusion detection[C]// *Proc. of the World Congress on Engineering*, 2007: 671－675.

[15] Tao Q, Wu G W, Wang F Y. Posterior probability support vector machines for unbalanced data[J]. *IEEE Trans. on Neural Networks*, 2005, 16(6): 1561－1573.

[16] KDD CUP 1999 Data [EB/OL]. [2011－07－16]. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.