# 基于SVM的网络入侵检测集成学习算法

谭爱平 陈 浩 吴伯桥
(湖南大学信息科学与工程学院 长沙 410082)

**摘 要** 互联网络中,计算机和设备随时受到恶意入侵的威胁,严重影响了网络的安全性。入侵行为升级快、隐蔽性强、随机性高,传统方法难以有效防范。针对这一问题,提出一种基于SVM的网络入侵检测集成学习算法,该算法利用SVM建立入侵检测基学习器,采用AdaBoost集成学习方法对基学习器迭代训练,生成最终的入侵检测模型,仿真实验表明了该算法的有效性。

**关键词** 安全,集成学习,入侵检测,AdaBoost,SVM
**中图法分类号** TP273 **文献标识码** A

## Network Intrusion Intelligent Detection Algorithm Based on AdaBoost

TAN Ai-ping CHEN Hao WU Bo-qiao
(School of Information Science and Engineering, Hunan University, Changsha 410082, China)

**Abstract** In the Internet, computers and equipment are threaded by malicious intrusion, and the safety of network is seriously affected. Intrusion behavior has features of upgraded fast, strong concealment, random characteristics, so the traditional methods are difficult to prevent this problem effectively. In this paper, a network intrusion intelligent detection algorithm based on AdaBoost was presented. The SVM is used to build the learning-module of intrusion detection. The AdaBoost is used for training these learning-modules, and generating the final the intrusion detection model. The simulation results show the effectiveness of the algorithm.

**Keywords** Security, Integrated learning, Intrusion detection, AdaBoost, SVM

## 1 引言

随着网络技术和社会经济的不断发展,人们在享用互联网和计算机技术便利的同时,也受到其所带来的安全威胁。防火墙作为传统的网络安全技术,对不断升级的网络入侵手段难以形成有效保护[1,2]。近年来作为主动保护策略,入侵检测日益受到国内外专家学者的重视,如何针对互联网现有入侵安全隐患,研究一种行之有效的入侵检测算法对互联网和经济的持续发展具有重要的意义。

传统的入侵检测方法主要分为异常检测、误用检测。异常检测多采用专家经验和推理的方式,具有代表性的算法有统计异常检测、贝叶斯推理异常检测等[3,4]。这类入侵检测方法对相对规律较强的入侵行为具有较好的检测效果,但对当今互联网不断升级的高技术入侵手段并不适用。误用检测能够将专家系统、预测算法有效结合,典型的算法有专家系统入侵检测、条件概率误用检测等。专家系统中规则难以快速更新,人为因素较大,效果不好;而条件概率误用检测方法,往往对时间相关性依赖较强,适用范围较小。近年来,将数据挖掘技术和智能算法引入入侵检测中,提升了检测的准确性,然而由于其需要大量完备的审计数据集作为支撑,难以应对目前网络入侵技术隐蔽性强、更新快的现状。

集成学习是一种建立在统计学习理论基础之上的机器学习方法,能够对学习算法泛化能力进行极大提升,在训练集样本有限的条件下,能够保证测试集对独立,保持较小的误差。在入侵检测中引入集成学习方法,可在先验知识不足的情况下仍保证有较好的分类正确率,从而使得入侵检测系统具有较好的检测性能。因此本文将集成学习算法和支持向量机相结合,提出一种基于SVM的网络入侵检测集成学习算法。

## 2 算法原理及系统结构

入侵检测算法的基本目标是:通过收集和分析网络数据,检测系统中可能存在的违反安全策略的行为和被攻击的迹象。通过积极主动的安全防护,在网络系统受到危害之前将入侵拦截并响应。

针对这一目标,本文提出一种基于SVM的网络入侵检测集成学习算法,其结构如图1所示。数据采集模块对来自各种信息源的实时网络数据流进行收集。为了保证入侵检测的相应速度和检测精确性,采集到的数据由检测引擎进行入侵检测处理。检测引擎由特征选择和检测模型两部分组成,对采集到的数据进行特征选择,降低数据维度;利用分类算法

到稿日期:2013-04-08 返修日期:2013-07-21 本文受国家自然科学基金项目(61272190)资助。
谭爱平(1979—),男,硕士,主要研究方向为网络工程、网络安全;陈 浩(1977—),男,博士,副教授,主要研究方向为网络安全和信息安全;吴伯桥(1979—),男,硕士,主要研究方向为网络系统集成、网络安全。

对特征子集进行学习，建立分类模型。

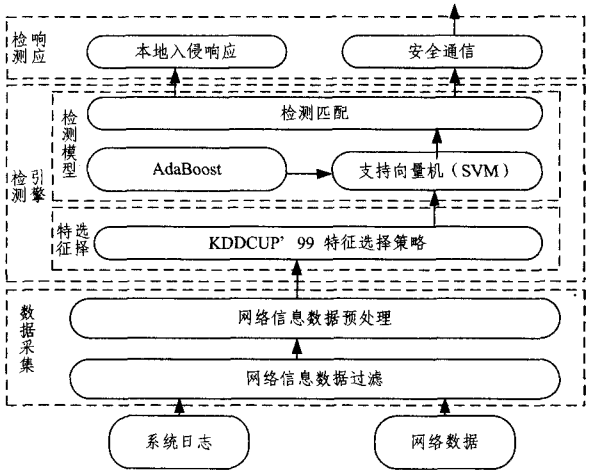


图1 入侵检测算法原理

分类算法是本文的重点，其思路为：将支持向量机(support vector machine，简称 SVM)和 AdaBoost 相结合，利用 SVM 对网络特征数据进行学习，获得入侵检测模型，作为基学习器，同时为了解决 SVM 对小样本随机问题精度不高的问题，引入 AdaBoost 集成学习算法，对 SVM 建立的入侵检测基学习器进行迭代优化，提升算法精确性。

## 3 基于 SVM 的入侵检测学习器

由 Vapnik 等学者提出的 SVM 算法既能有效地处理非线性数据，又能限制过学习，同时具有严格的理论基础和数学基础，不存在局部最小问题，对于网络入侵检测这类小样本学习应用具有很强的泛化能力，对样本数量的依赖性弱[5,6]。标准的支持向量机算法是一个凸二次优化问题，总可以找到全局最优点。但是当训练样本增多时，由于其约束过多，将导致训练时间和内存需求大大增加，这成为了 SVM 在实际应用中的瓶颈。为提高 SVM 的训练效率，Suyken 改变了标准 SVM 的约束条件和风险函数，提出了最小二乘支持向量机(LS-SVM)[7]，LS-SVM 的训练只要求解一个线性方程组，使 SVM 易于实现，并极大地提高了 SVM 的训练效率，因此本文采用 LS-SVM 方法作为网络入侵检测的基学习器。

在实际网络环境中，各个网络节点都会接受大量的网络数据，这些数据中，只有很小一部分信息表征了入侵行为，为了减少数据处理量，本文在 KDDCUP'99 特征选择策略的基础上进行改进，给出如下评价函数：

$$w_n(d)=\frac{P_{ik}\times tf_{ik}\times\log(N/n_k+0.001)}{\sqrt{\sum_{k=1}^{n}(P_{ik}\times tf_{ik})^2\times\log(N/n_k+0.001)}}\times(1-1/L) \tag{1}$$

式中，$P_{ik}$ 表示网络数据 $t_k$ 在网络数集 $d$ 中的数据来源权重；$tf_{ik}$ 表示某一数据 $t_k$ 在网络数集 $d$ 中的出现频数；$N$ 为网络数集 $d$ 中的数据量；$n_k$ 表示含有特定端口号的数据频数；$L$ 为 $t_k$ 的长度。选取权重较大的报文作为训练样本。

对于训练样本集 $(x_i,y_i)$(其中，$i=1,2,\cdots,n$；$x_i\in R^n$，为输入变量；$y_i\in R$，为对应的输出值)，样本包含 9 个基本特征、13 个内容特征、9 个两秒钟内的流量特征和 10 个主机流量特征。SVM 回归理论的基本思想是寻找一个输入空间到输出空间的非线性映射 $\varphi$，通过这个非线性映射[8,9]，将数据 $x$ 映射到一个高维特征空间 $F$，并在特征空间中用下列估计函数进行线性回归，如式(2)所示：

$$f(x)=[\omega\times\varphi(x)]+b,\phi:R^n\to F,\omega\in F \tag{2}$$

式中，$b$ 为阈值。函数逼近问题等价于式(3)。

$$R_{reg}(f)=R_{emp}(f)+\lambda\|\omega\|^2=\sum_{i=1}^{s}C(e_i)+\lambda\|\omega\|^2 \tag{3}$$

式中，$R_{reg}(f)$为目标函数；$s$ 为样本数量；$\lambda$ 为调整常数；$C$ 为错误惩罚因子；$\|\omega\|^2$ 反映 $f$ 在高维空间平坦的复杂性。

考虑到线性 $\varepsilon$ 不敏感损失函数具有较好的稀疏性，可以得到以下损失函数，如式(4)：

$$|y-f(x)|_\varepsilon=\max\{0,|y-f(x)-\varepsilon\} \tag{4}$$

经验风险函数如式(5)：

$$R^{\varepsilon}_{emp}(f)=\frac{1}{n}\sum_{i=1}^{n}|y-f(x)|_\varepsilon \tag{5}$$

根据统计学理论，SVM 通过对以下目标函数极小化确定回归函数如式(6)：

$$\begin{cases}\min\{\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{n}(\xi_i^*+\xi_i)\}\\ y_i-\omega\cdot\phi(x)-b\leqslant\varepsilon+\xi_i^*\\ \omega\cdot\phi(x)+b-y_i\leqslant\varepsilon+\xi_i\\ \xi_i,\xi_i^*\geqslant 0\end{cases} \tag{6}$$

式中，$C$ 为用于平衡模型复杂项和训练误差项的权重参数；$\xi_i,\xi_i^*$ 为松弛因子；$\varepsilon$ 为不敏感损失函数。该问题可转化为式(7)所示的对偶问题：

$$\begin{cases}\max=-\frac{1}{2}\sum_{i,j=1}^{n}(a_i^*-a_i)(a_j^*-a_j)K(x_i,y_j)x-\\ \qquad\frac{1}{2}\sum_{i,j=1}^{n}(a_i^*-a_i)(a_j^*-a_j)K(x_i,y_j)+\\ \qquad\sum_{i}^{l}a_i^*(y_i-\varepsilon)-\sum_{i=1}^{n}a_i(y_i-\varepsilon)\}\\ \sum_{i=1}^{n}a_i=\sum_{i=1}^{n}a_i^*\\ 0\leqslant a_i^*\leqslant C\\ 0\leqslant a_i\leqslant C\end{cases} \tag{7}$$

利用拉格朗日乘子法和核技术，最小二乘支持向量机则可转化为求解如下线性方程组：

$$\begin{bmatrix}0 & \vec{1}^{\mathrm{T}}\\ \vec{1} & \Omega+C^{-1}I\end{bmatrix}\begin{bmatrix}b\\ a\end{bmatrix}=\begin{bmatrix}0\\ Y\end{bmatrix} \tag{8}$$

式中，$\Omega_{i,j}=k(x_i,x_j)=\psi(x_i)\psi^{\mathrm{T}}(x_j)$，$i,j=1,2,\cdots,l$，$Y=[y_1,y_2,\cdots,y_l]$，$a=[a_1,a_2,\cdots,a_l]$，$b=[b_1,b_2,\cdots,b_l]$，$\vec{1}=[1,1,\cdots 1]$。

核函数 $k(x_i,x_j)=\psi(x_i)\psi^{\mathrm{T}}(x_j)$，为满足 Mercer 条件的任意对称函数，核函数的选择需要一定的先验知识，目前还没有一般性的结论，Scholkopf 等就核函数的选择和构造作了讨论。对于 SVM 的黑箱模型构建，最主要的是核函数的选取，常见的核函数有线性函数、多项式函数、径向基函数、多层感知器函数。本文采用径向基函数，如式(9)：

$$K(x_i,x_j)=\exp[\frac{\|x_i-x_j\|}{\sigma^2}] \tag{9}$$

求解上述问题，可得到 SVM 回归函数，如式(10)：

$$f(x)=\sum_i^n(a_i-a_i^*)K(x_i,x_j)+b \tag{10}$$

从而得到网络入侵检测的基学习器。对于未知属类的网络数据向量 $x$，可以采用式(11)所示的线性判决函数进行分类。

$$g(x)=\mathrm{sgn}(\sum_i^n(a_i-a_i^*)K(x_i,x_j)+b) \tag{11}$$

## 4 AdaBoost 集成学习算法

网络中，节点受到入侵威胁的因素很多，导致入侵过程的发生时间、特征信息呈现一定的弱随机性，单纯的 SVM 算法虽然具有一定的小样本泛化能力，但对于入侵检测问题仍然精度不高。AdaBoost 是典型的集成学习方法，能够将多个精度相对较低的弱学习算法综合优化[10,11]，训练出精度较高的强学习算法，提升预测精度。本文将 AdaBoost 用于入侵检测 SVM 基学习器的训练，原理是：用 SVM 算法生成一系列的基学习器，每个基学习器的训练依赖于上一次基学习器的学习结果[12,13]。基学习器在训练集上的错误率用于调整训练样本的概率分布，通过单个基分类器加权建立最终的入侵检测模型。

基于 AdaBoost 的模型集成学习算法原理如图 2 所示，首先根据 SVM 的预测误差计算样本权重，初始化时，各个样本子集的样本权重相同；然后利用加权后的样本进行 SVM 训练，得到该次入侵检测模型，同时更新该模型的模型权重；接着根据迭代次数或者模型精度是否达到设定值来判断是否结束迭代，如果迭代结束，则根据模型权重和各次入侵检测模型生成最终的入侵检测模型，否则重新计入样本权重计算，开始新一轮的迭代。

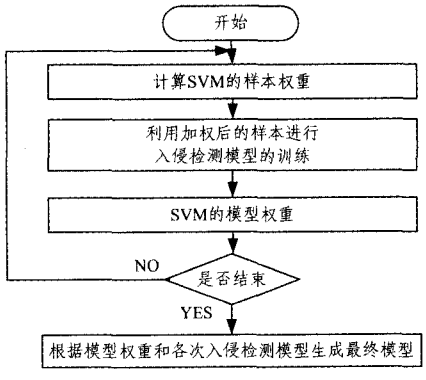


图 2 基于 AdaBoost 的集成学习算法原理

本文将 AdaBoost 方法用于入侵检测的 SVM 基学习器训练中，具体实现步骤如下：

Step1 设 $m$ 个初始学习的网络数据特征样本集 $s=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$，其中 $x_m$ 为一向量，是网络数据特征训练样本，$y_m$ 为对于入侵检测问题的分类结果，各个样本初始权重 $d_1,d_2,\cdots,d_m$ 均设置为 $\frac{1}{m}$，设置 AdaBoost 算法最大迭代次数为 $T$，并初始化当前迭代次数 $t=1$。

Step2 针对 $m$ 个训练集，利用算法对 SVM 的连接权重值进行优化选择，得到最优 SVM 的连接权重值。

Step3 利用优化后的 SVM 对 $m$ 个训练集分别进行训练，获得第 $t$ 次的入侵检测模型 $h_t$。

Step4 记录本次入侵检测模型 $h_t$，计算并保存第 $t$ 次入侵检测模型 $h_t$ 的权重 $\omega_t$，根据 Step3 得到的入侵检测模型对 $m$ 个训练集的预测误差绝对值和小于设定值，或达到最大迭代次数，算法结束，跳出迭代进入 Step6；否则进入 Step5。

Step5 根据入侵检测模型对 $m$ 个训练集的预测误差绝对值和，更新 $m$ 个训练的权重 $d_1,d_2,\cdots,d_m$，生成新的样本，返回 Step2，进行迭代。

Step6 得到最终的预测模型 $h=\sum_{t=1}^{T}\omega_t h_t$。

影响 AdaBoost 集成学习效果的因素主要有两个，一是每一轮循环中训练集上的样本权重如何分布；二是许多个规则如何集成为一个有效的预测规则。这两点分别通过样本权重和模型权重来体现。

### 4.1 样本权重计算

通过对样本权重值的调节，能够有效地降低错误样本对入侵检测模型的贡献，提升正确样本的影响。样本权重值分为计算和归一化处理两个步骤，其中权重值采用预测误差绝对值进行衡量，其方法如式(12)：

$$\begin{cases}E_t=\sum_{k=1}^{N}(d_t(k)(h_t(k)-y_t(k)))^2\\ \beta_t=\dfrac{E_t}{1-E_t}\\ d'_{t+1}(k)=d_t(k)\cdot\beta_t\exp(1-\dfrac{h_t(k)-y_t(k)}{\max\limits_{t=1}^{N}|h_t(k)-y_t(k)|})\end{cases} \tag{12}$$

式中，$E_t$ 表示第 $t$ 次迭代得到的入侵检测模型对各个训练样本的加权方差和，$\beta_t$ 为调节系数，$d'_{t+1}(k)$ 为新样本权重。

各样本的权重值综合必须为 1，因此必须进行归一化处理，其方法如式(13)：

$$d_{t+1}(k)=\frac{d'_{t+1(k)}}{\sum_{k=1}^{N}d'_{t+1(k)}} \tag{13}$$

### 4.2 模型权重计算

入侵检测模型的权重 $\omega_t$ 的计算，直接影响了最终的预测模型的输出。为了提升误差较小的入侵检测模型 $h_t$ 在最终的模型中的影响权重，本文利用预测误差绝对值来衡量这一权重值，其方法如式(14)：

$$\begin{cases}E_t=\sum_{k=1}^{N}(d_t(k)(h_t(k)-y_t(k)))^2\\ \beta_t=\dfrac{E_t}{1-E_t}\\ \omega_t=\dfrac{1}{2}\cdot\ln(\dfrac{1}{\beta_t})\end{cases} \tag{14}$$

式中，$E_t$ 表示第 $t$ 次迭代得到的入侵检测模型对各个训练样本的加权方差和，$\beta_t$ 为调节系数，调节系数的选取有多种方式，为了保证最终的预测模型能够稳定，本文采用了上述方式，$\omega_t$ 为最终输出的第 $t$ 次迭代得到的入侵检测模型对最终预测模型的影响权重值。

## 5 仿真与实验

为了验证算法的有效性，按照本文提出的入侵检测算法进行计算机仿真实验，实现平台采用 Windows XP，Matlab 7 语言编程环境，3.0GHz CPU 时钟频率，4G 内存。

在评估入侵检测系统性能方面，大多数专家学者普遍采用 DARPA'98 数据进行验证，为保证仿真实验的权威性，本文亦采用该数据集对算法进行评估，数据集被分为训练集(Training Set)(包含有 500 万个连接)和测试集(Test Set)(包含有 311029 个连接数据)。测试集中包含有训练集中没有出现过的攻击。

本文从样本中抽取 41 维 29313 个数据作为训练集，其中包含 6059 个"Normal"、3866 个"Neptune"、516 个"Portsweep"、177 个"Satan"，11 个"Buffer_overflow 和 2183 个"Guess-password"；从 TestData 中抽取 124970 个数据划分为 5 个测试集，分别以单纯 BP 神经网络、单纯 SVN，以及本文提出的基于 SVM 的网络入侵检测集成学习算法进行实验，测试结果如表 1－表 3 所列。其中，1 表示"Normal"、2 表示"Neptune"、3 表示"Portsweep"、4 表示"Satan"、5 表示"Buffer_overflow、6 表示"Guess-password"。

从表 1 中可以看出，采用单纯 BP 神经网络算法进行入侵检测，由于训练集中"Satan"和 "Buffer_overflow"样本较少，且网络数据具有一定的随机性，测试精度略好于其他入侵检测项，误差相对较大，若样本较大，效果有所提升，但精度仍然较低。

表 1 单纯 BP 神经网络入侵检测仿真结果

| 测试集 \ 准确率 | 1(%) | 2(%) | 3(%) | 4(%) | 5(%) | 6(%) |
|---|---|---|---|---|---|---|
| 测试集 1 | 85.2 | 78.5 | 71.3 | 65.6 | 61.2 | 75.4 |
| 测试集 2 | 81.5 | 74.6 | 72.5 | 68.3 | 53.5 | 73.6 |
| 测试集 3 | 81.2 | 75.3 | 75.3 | 72.2 | 48.4 | 75.1 |
| 测试集 4 | 84.3 | 74.3 | 74.6 | 64.6 | 58.7 | 74.2 |
| 测试集 5 | 85.6 | 76.8 | 72.3 | 66.6 | 57.3 | 72.3 |

将表 2 结果与表 1 结果对比可以看出，SVM 由于对于小样本集的泛化能力略好于神经网络，因此对于训练集中"Satan"和 "Buffer_overflow"测试项，测试精度略好于神经网络入侵检测方法，而对其样本集较大的测试项，精度相差不大，检测精度总体略有提升，但仍然较低。

表 2 单纯 SVM 入侵检测仿真结果

| 测试集 \ 准确率 | 1(%) | 2(%) | 3(%) | 4(%) | 5(%) | 6(%) |
|---|---|---|---|---|---|---|
| 测试集 1 | 88.4 | 77.8 | 73.5 | 69.7 | 72.2 | 74.9 |
| 测试集 2 | 82.5 | 77.6 | 75.5 | 75.3 | 75.5 | 78.6 |
| 测试集 3 | 83.2 | 76.3 | 74.3 | 78.2 | 62.4 | 76.1 |
| 测试集 4 | 85.3 | 72.3 | 72.6 | 72.6 | 68.7 | 72.2 |
| 测试集 5 | 88.6 | 75.8 | 75.3 | 75.6 | 69.3 | 73.3 |

从表 3 中可以看出，由于采用了 AdaBoost 算法对 SVM 模型进行了迭代修正，大大降低了随机样本对模型的影响，也大大增加了 SVM 的泛化能力，相对于单纯的 BP 神经网络算法和 SVM 算法，最终的入侵检测模型更加贴近真实的网络入侵样本，减小了小样本集导致的模型精度大幅下降的问题；对"Satan"和 "Buffer_overflow"类型的入侵，检测精度仍然有所保证，同时模型的整体检测精度也有较大的提升。

表 3 基于 SVM 的网络入侵检测集成学习算法仿真结果

| 测试集 \ 准确率 | 1(%) | 2(%) | 3(%) | 4(%) | 5(%) | 6(%) |
|---|---|---|---|---|---|---|
| 测试集 1 | 98.2 | 95.5 | 96.4 | 89.6 | 86.3 | 96.4 |
| 测试集 2 | 95.5 | 93.6 | 95.4 | 92.3 | 85.5 | 95.6 |
| 测试集 3 | 97.2 | 96.3 | 93.2 | 87.2 | 87.8 | 97.1 |
| 测试集 4 | 98.3 | 95.3 | 92.3 | 85.6 | 83.6 | 93.2 |
| 测试集 5 | 97.6 | 93.8 | 93.1 | 88.6 | 89.2 | 93.3 |

**结束语** 针对网络中入侵行为攻击隐蔽性强、特征变化快、随机性高的特点，本文提出一种基于 SVM 的网络入侵检测集成学习算法。利用 SVM 对权重较高的特征报文进行学习，建立入侵检测基学习器，采用 AdaBoost 集成学习方法对 SVM 基学习器进行迭代训练，生成最终的入侵检测模型。通过对比实验不难看出，其由于在 SVM 的基础上增加了 AdaBoost 的集成学习，解决了 SVM 对入侵检测这类样本小、随机高的对象的拟合精度不高的问题，检测率有了较大的提升，可以推广在现有网络中使用。

## 参 考 文 献


[1] Amor N, Benferhat S, Elouedi Z. Naive Bayes VS Decision Trees in Intrusion Detection System[C] // Proceedings of the 2004 ACM Symposium on Applied Computing. 2004: 420-424

[2] Sebring M, Shellhouse M, Hanna E. Expert systems in intrusion detection: a case study[C] // Proceedings of the 11th National Computer Security Conference. Baltimore, Maryland, 2006: 74-81

[3] Fuchsberger A. Intrusion Detection Systems and Intrusion Prevention Systems[J]. Information Security Technical Report, 2005, 34(10): 134-139

[4] Hu Y, Perrig A, Johnson D. Wormhole attacks in wireless networks[J]. IEEE Journal on Selected Areas in Communications, 2006, 24(2): 370-380

[5] 王振树，李林川，牛丽. 基于贝叶斯证据框架的 SVM 负荷建模[J]. 电工技术学报，2009，24(8)：83-86

[6] Davy M, Desobry F, Arthur G. An online support vector machine for abnormal events detection[J]. Signal Processing, 2006, 86(8): 2009-2025

[7] Bo Cui-mei. Research on the modeling method based on eliding time window for support vector machine soft-sensing [J]. Automatic Instrument, 2006, 27(1): 45-51

[8] 李良敏，温广瑞，王生昌. 基于遗传算法的回归型 SVM 参数选择法[J]. 计算机工程与应用，2008，44(7)：23-26

[9] 颜根廷，李传江，马广富. 基于混合遗传算法的 SVM 参数选择[J]. 哈尔滨工业大学学报，2008，40(5)：134-138

[10] 陈爱斌，夏利民，赵桂敏. 基于 Boosting 方法的人脸检测[J]. 计算机工程与应用，2004，27(3)：48-52

[11] 吴飞，庄永真，潘红. 基于分形布朗运动和 AdaBoosting 的多类音频例子识别[J]. 计算机研究与发展，2003，18(7)：62-65

[12] Diao Li-li, Hu Ke-yun, Lu Yu-chang. Improved Stumps Combined by boosting for Text Categorization[J]. Journal of Software, 2002, 25(8): 51-53

[13] 夏利民，戴汝为. 基于 Boosting 模糊分类的滚动轴承故障诊断[J]. 模式识别与人工智能，2003，18(3)：72-75