文章编号: 1001-0920 (2015) 01-0081-05 DOI: 10.13195/j.kzyjc.2013.1562

基于 MultiBoost 的集成支持向量机分类方法及其应用

吕 锋, 李 翔, 杜文霞

(河北师范大学 物理科学与信息工程学院，石家庄 050024)

摘 要: 针对网络故障诊断中的模式识别问题, 提出一种基于多重提升 (MultiBoost) 的优化支持向量机集成学习方法. 首先, 利用自适应的荷尔蒙调节遗传算法 (HMGA), 对支持向量机基分类器进行建模参数优化; 然后, 通过构建 MultiBoost 集成学习方法将多个基分类器集成, 建立以多分类器优化集成为核心的故障诊断系统. 实验结果表明, 所提出的方法在网络故障诊断中, 迭代次数少、建模时间短, 并且能够明显提高故障分类的准确率.

关键词: 支持向量机；荷尔蒙调节遗传算法；多分类器集成；网络故障诊断

中图分类号: TP29 **文献标志码:** A

MultiBoost with SVM-based ensemble classification method and application

LÜ Feng, *LI Xiang*, *DU Wen-xia*

(School of Phycisis Science and Information Engineering，Hebei Normal University，Shijiazhuang 050024，China. Correspondent：LI Xiang，E-mail：lixiang_hebtu@163.com)

Abstract: For pattern recognition in the network fault diagnosis, an optimal SVM ensemble learning method based on MultiBoost is proposed. Firstly, the parameters of SVM-base-classifier are optimized by using the adaptive hormone modulation genetic algorithm(HMGA). Then, multi-base-classifiers are integrated by using the MultiBoost algorithm. Finally, with multiple ensemble optimal classifiers as the core, the fault diagnosis system is established. Simulation results show that the proposed method can not only reduce the number of iteration and lower the computing cost, but also improve the fault diagnosis accuracy of the network fault diagnosis system.

Keywords: SVM；hormone modulation GA；multi-classifier ensembles；network fault diagnosis

0 引 言

分类是机器学习与模式识别领域的基本问题之一, 支持向量机 (SVM) 以其出色的非线性和泛化性能在该问题中获得了良好的效果, 并成功应用于故障诊断领域[1]. 网络异常检测与诊断其实质是一类模式识别问题[2], 分类器的构造在此类问题中至关重要. SVM 及其改进算法在网络故障诊断中得到较为广泛的应用, 主要有: 1) 在研究训练数据对分类器影响方面, 文献 [2] 提出双重支持向量机以减少训练样本个数, 文献 [3] 提出了采用粗糙支持向量机来减小训练数据属性个数的方法, 文献 [4] 研究了如何降低数据集不平衡样本对诊断结果的影响; 2) 在研究 SVM 二次规划求解方面, 文献 [5] 提出双子支持向量机以加快二次规划求解问题的求解速度, 文献 [6] 提出近轴支持向量机以避免二次规划求解问题.

使用 SVM 时需要设定由软间隔分类所需的规则化参数 C 和高斯核函数的高斯宽度 σ. 通常, 参数 (C, σ) 的选取直接影响分类器的分类效果. 本文将优化理论应用于 SVM 参数寻优, 将标准遗传算法改进为一种自适应的荷尔蒙调节遗传算法 (HMGA), 以优化参数寻优过程.

基于集成学习方法, 可将多个分类器进行集成, 利用组合决策分类的结果提高分类的准确率. 多重提升 (MB) 算法将经典算法自适应提升算法 (AB)[7] 与装袋 (Bagging)[8] 算法的改进算法 Wagging[9] 进行统一构造, 可利用 AB 减小偏差的能力较强来降低分类误差. 此外, Bagging 和 Wagging 减小偏差的能力较弱, 却有较强的减小方差的能力, MB 算法将二者结合后, 在有

收稿日期: 2013-11-08；**修回日期:** 2014-04-19.

基金项目: 国家自然科学基金项目(60974063, 61175059)；河北省自然科学基金项目(F2014205115)；河北省高等学校科学技术研究项目(Q2012053).

作者简介: 吕锋(1958−), 女, 教授, 从事信息处理及故障检测等研究；李翔(1987−), 男, 硕士生, 从事模式分类与智能系统的研究.

效减小分类模型偏差的同时也有效减小了方差, 从而降低了分类误差[10].

本文提出一种基于多重提升的荷尔蒙遗传SVM (MBHMSVM) 集成学习方法. 将其应用于网络故障诊断系统中分类器的设计, 可提高故障分类准确率. MBHMSVM将经过HMGA参数优化后的SVM分类器作为基分类器, 再利用MB算法组合多个基分类器, 遵循共同决策的原则进行组合分类.

1 基于HMGA的优化参数基分类器

1.1 SVM基分类器

设有样本集$\{(x_i, y_i)\}, i \in \{1, 2, \cdots, N\}$. 其中: N为样本总数, $y_i = \{1, -1\}$, $x_i \in \boldsymbol{R}^p \subset \boldsymbol{R}$. 设该样本集能被超平面分类, 其约束优化方程为

$$\begin{aligned} &\min\ \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}\xi_i;\\ &\text{s.t. } y_i(\boldsymbol{w}^{\mathrm{T}}x_i + b) \geqslant 1 - \xi_i,\\ &\qquad \xi_i \geqslant 0,\ i = 1, 2, \cdots, N. \end{aligned} \tag{1}$$

其中: $\boldsymbol{w}$为权向量, 垂直于超平面; b为超平面的偏置; ξ_i为松弛变量, 用于计量数据集x_i中错分样本的错分程度; C为规则化参数, 用于对非零的ξ_i值进行错分惩罚.

式(1)的对偶问题为

$$\begin{aligned} &\min\ \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i \cdot x_j);\\ &\text{s.t. } 0 \leqslant \alpha_i \leqslant \sum_{i=1}^{N}\alpha_i y_i,\ i = 1, 2, \cdots, N. \end{aligned} \tag{2}$$

其中: $\{\alpha_i\}_{i=1}^{N}$为Lagrange乘子, $K(x_i \cdot x_j)$为核函数. 得到最优值α_j后, 对于未知样本$\boldsymbol{z}$的类别, 可用下式的符号函数表示:

$$G(\boldsymbol{z}) = \sum_{j=1}^{N}\alpha_j y_j K(\boldsymbol{z}, x_j). \tag{3}$$

1.2 基分类器参数优化

HMGA参照荷尔蒙释放机制的仿真模型[11]对传统遗传算法进行了改进, 应用HMGA算法进行SVM中(C, σ)参数寻优. 每个染色体为实数编码机制下的(C, σ)值, 目标函数为训练集在交叉验证意义下的分类准确率, 按此机制进行遗传进化的计算.

按照荷尔蒙腺释放荷尔蒙遵从的一般规律, 一种荷尔蒙x的释放速度S_x受另一种荷尔蒙y的影响为

$$S_x = \alpha F(C_y) + S_{x_0}. \tag{4}$$

其中: S_{x_0}是x的原始释放速度, α为常量.

y的荷尔蒙浓度通常满足下式:

$$F(C_y) = \frac{S_{x_0}\cdot(S_{y_{\mathrm{av}}})^n}{(S_{y_{\max}} - S_{y_{\min}})^n + (S_{y_{\mathrm{av}}})^n}. \tag{5}$$

其中: $S_{y_{\mathrm{av}}}$、$S_{y_{\max}}$和$S_{y_{\min}}$分别代表y释放速度的平均值、最大值和最小值, n为常数.

传统遗传算法中的交叉概率p_{c}和变异概率p_{m}通常为经验给定的常数. 受荷尔蒙调节机制启发, 可将p_{c}、p_{m}设定为自适应取值, 令其在进化过程中每一代的取值均受到前一代的影响.

1) 交叉概率的改进. 为加快进化速度并扩大搜索范围, 改进自适应交叉概率为

$$P_{\mathrm{c}} = P_{\mathrm{c}}^0\Big(1 + \alpha\frac{(g_{\mathrm{av}})^{n_{\mathrm{c}}}}{(g_{\max} - g_{\min})^{n_{\mathrm{c}}} + (g_{\mathrm{av}})^{n_{\mathrm{c}}}}\Big). \tag{6}$$

其中: P_{c}^0为初始交叉概率; g_{av}、$g_{\min}$和$g_{\max}$分别为每一代中平均适应度、最低适应度和最高适应度; α和n_{c}为参数. 可见, 交叉概率与平均适应度之间存在着正相关的变化关系.

2) 变异概率的改进. 为加快进化速度并保持种群的多样性, 改进自适应变异概率为

$$P_{\mathrm{m}} = P_{\mathrm{m}}^0\Big(1 + \beta\frac{(g_{\mathrm{av}})^{n_{\mathrm{m}}}}{(g_{\max} - g_{\min})^{n_{\mathrm{m}}} + (g_{\mathrm{av}})^{n_{\mathrm{m}}}}\Big). \tag{7}$$

其中: P_{m}^0为初始变异概率; β和n_{m}为参数. 下一代种群中新个体的生成将按照改进后的交叉概率和变异概率进行.

2 基于MultiBoost的多分类器集成方法

2.1 Wagging算法

在Bagging算法中, 设数据集为$S = \{(x_i, y_i), x_i \in X, y_i \in Y\}$. 若将$T$个基分类器$(C_t, t = 1, 2, \cdots, T)$进行集成, 使用自引导随机抽样(Bootstrap)得到的分类结果以投票的方式按下式进行决策组合:

$$C(x) = \arg\max_{y\in Y}\sum_{i=1}^{T}1(C_t(x) = y), \tag{8}$$

其中$1(\cdot)$是指示函数.

不同于Bagging的Bootstrap取样方式, Wagging按照连续泊松分布方式为每个训练样本指定权值. 个体由连续泊松分布生成权值的计算方式为

$$\mathrm{Poisson}() = -\log\frac{\mathrm{Random}(1, 2, \cdots, 999)}{1\,000}. \tag{9}$$

2.2 MultiBoost算法

MB算法在T个基分类器集成中, 不同于AB方法直接进行T轮迭代, 而是定义了若干个子决策组和一个判断子决策组迭代次数的标志变量I_i $(i = 1, 2, \cdots, \infty)$. 第$i$个子决策组包含$I_i$个由AB算法构成的基分类器, 子决策组间使用Wagging策略进行组合. I_i的取值通常遵循如下规则:

$$\begin{cases} n = \lfloor\sqrt{T}\rfloor;\\ I_i = \lceil i \times T/n\rceil,\ i = 1, 2, \cdots, n-1;\\ I_i = T,\ i = n, n+1, \cdots, \infty. \end{cases} \tag{10}$$

无论子决策组每轮迭代产生的误差是否过高或过低, 下一个子决策组都将进行迭代且迭代终止数目

增加. 若最后一个子决策组迭代终止后仍存在分类误差过高或过低的现象, 算法将附加一个或多个子决策组, 直至满足基分类器对误差的要求. 算法开始时, 将数据集 $S=\{(x_i,y_i),x_i\in X,y_i\in Y\}$ 中每个样本的权重均设置为 1, 形成数据集 S'. 当每个子决策组迭代终止时, 对 S' 中的数据重新进行服从连续泊松分布的权重设置且进行归一化处理, 再进行下一个子决策组的迭代. 在每次基分类器迭代中, 基分类器误差为 $\epsilon_t\,(t=1,2,\cdots,T)$. 误差可根据下式进行计算:

$$\epsilon_t=\frac{\sum\limits_{x_j\in S',C_t(x_j)\neq y_j}\mathrm{weight}(x_j)}{m},\tag{11}$$

基分类器对最终组合分类模型的输出权重设为 β_t, $t=1,2,\cdots,T$.

根据 ϵ_t 每次取值范围的不同, 分为 3 种情况:

1) 若 $\epsilon_t>0.5$, 则直接将该分类器舍弃, 并再对训练集进行服从连续泊松分布的权重设置且进行归一化处理, 进入下一个子决策组, 直至 $\epsilon_t\leqslant 0.5$.

2) 若 $\epsilon_t=0$, 则设置 $\beta_t=10^{-10}$, 并再对训练集进行服从连续泊松分布的权重设置且进行归一化处理, 进入下一个子决策组.

3) 若 $0<\epsilon_t\leqslant 0.5$, 则设置 $\beta_t=\epsilon_t/(1-\epsilon_t)$, 对于 S' 中的每个样本, 错分样本的权重除以 $2\epsilon_t$; 正分样本的权重乘以 $2(1-\epsilon_t)$, 但最小权重为 10^{-8}. 最终的分类函数为

$$C^*(x)=\arg\max_{y\in Y}\sum_{t:C_t(x)=y}\log\frac{1}{\beta_t}.\tag{12}$$

3　基于 MBHMSVM 的故障诊断方法

3.1　MBHMSVM 故障诊断框架

基于MBHMSVM的故障系统框架如图 1 所示, 图 1 包括分类器核心模块部分和其他外围模块部分. 构建分类核心模块时, 将 SVM 选取为基分类器, 并应用 HMGA 进行参数优化. 历史样本库中的系统状态数据作为基分类器的训练集, 在每轮迭代中按照连续泊松分布的权值取样机制依次输入基分类器中, 再利用 MB 算法将多个基分类器集成, 以获得一个准确率较高的决策分类结果.

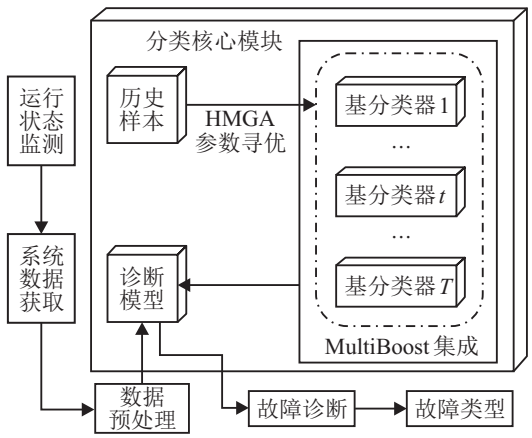


图 1　基于 MBHMSVM 的故障诊断系统框架

在故障诊断系统中实施, 还需建立系统外围模块部分. 系统状态监测模块对系统进行实时监视, 以一定时间间隔获取系统状态数据, 并按照要求进行预处理, 预处理后的数据将输入到分类器核心模块中进行故障检测与诊断. 若存在故障, 系统将自动识别故障类型, 为下一步维护决策提供依据.

3.2　MBHMSVM 故障分类算法流程

针对图 1 所示故障诊断系统框架中的分类核心模块, 依据前文部分理论方法, MBHMSVM 算法的描述如下.

输入: $S=\{(x_i,y_i),x_i\in X,y_i\in Y\}$ 为历史样本库中数据集; 整数 T 为基分类器 (单 SVM 分类器) 集成个数; 整数变量 I_i 为第 $i\,(i\geqslant 1)$ 个子决策组的终止标志变量; 辅助变量 $k=1$.

输出: 输出策略按式 (12) 计算获得.

Step 1: 利用 HMGA 方法在 S 的 n 分交叉验证意义下求得 SVM 基分类器建模最优参数 (C,σ).

Step 2: 利用最优参数 (C,σ) 建立单 SVM 故障分类模型.

Step 3: 设 S 中每个元素的初始权重均为 1, 记为 $\mathrm{weight}(x_j)=1$, 将这时的训练集定义为 S'.

Step 4: For $t=1,2,\cdots,T$:

Step 4.1: 根据式 (12) 计算 I_i 的值;

Step 4.2: 若 $I_k=t$, 则将 S' 中所有元素的权重随机重置为连续泊松分布并进行归一化处理, $k++$;

Step 4.3: 用所建立的单 SVM 分类器对 S' 进行故障分类, 表示为 $C_t=\mathrm{SVM}_{(C,\sigma)}(S')$;

Step 4.4: 根据诊断分类结果计算故障误识率, 记为 ϵ_t, 其计算方法按式 (11) 进行, 基分类器对最终组合分类模型的输出权重记为 β_t.

根据 ϵ_t 每次取值范围不同, 分为 3 种情况:

1) 若 $\epsilon_t>0.5$, 则将 S' 中元素权重随机重置为连续泊松分布并进行归一化处理, $k++$; 转向 Step 4.2;

2) 若 $\epsilon_t=0$, 则将 S' 中元素权重随机重置为连续泊松分布并进行归一化处理, $k++$;

3) 若 $0<\epsilon_t\leqslant 0.5$, 则令 $\beta_t=\dfrac{\epsilon_t}{1-\epsilon_t}$, 并对 S' 中元素进行权重更新, 分类正确样本 $\mathrm{weight}(x_j)=\mathrm{weight}(x_j)/2(1-\epsilon_t)$, 分类错误样本 $\mathrm{weight}(x_j)=\mathrm{weight}(x_j)/2\epsilon_t$, $\mathrm{weight}(x_j)$ 下限为 10^{-8}.

4　实验与讨论

4.1　实验数据集

本文选用 UCI 中网络故障诊断与安全领域经典

数据集KDD99进行实验. 该数据集中有正常(NOR)和拒绝服务攻击(DoS)、来自远程主机的未授权访问(R2L)、未授权的本地超级用户特权访问(U2R)、端口监视与扫描(PROBE)四大类异常状态的网络故障. 从KDD99中抽取4 500条记录形成数据集, 其中训练集2 505条, 测试集1 995条. 为模拟真实环境同时测试方法的鲁棒性, 向数据集中加入50 dB随机噪声, 然后对数据集进行归一化处理.

4.2 SVM基分类器参数寻优

(C,σ)的取值直接影响SVM的分类表现: 过高的参数C可能导致SVM的分类效果过于拟合, 过低的参数C可能影响SVM对线性不可分数据的分类性能. 在使用RBF高斯核函数$K(x_i,x_j)=\exp(-\|x_i-x_j\|^2/2\sigma^2)$时, σ^2的不同取值对构建出的高斯核函数自身变化趋势与输出值存在较大差异, 不同的σ参数值在用于构建核函数时, 性能表现差异很大.

为加快参数寻优速度, 克服机械枚举寻优法和实验寻优法的投机性等弊端, 采用HMGA算法进行SVM中(C,σ)参数寻优. 将(C,σ)以实数编码方式作为染色体, 以SVM分类器在训练集上采用5分交叉验证下的分类准确率作为目标函数, 进行各代遗传迭代计算. 初始参数设置为: 种群容量为20, 遗传进化代数为50, 初始交叉概率为0.7, 初始变异概率为0.01. 在算法执行过程中, 变异概率和交叉概率分别按照式(6)和(7)进行自适应更新, 式中参数设置为: $\alpha=0.3,\beta=0.2,n_c=n_m=2$. 运算50代的遗传计算过程如图2所示, 其中的曲线分别为每代种群的平均适应度和最佳适应度.

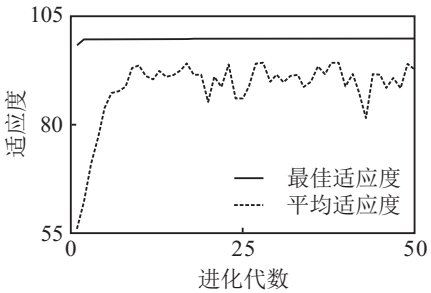


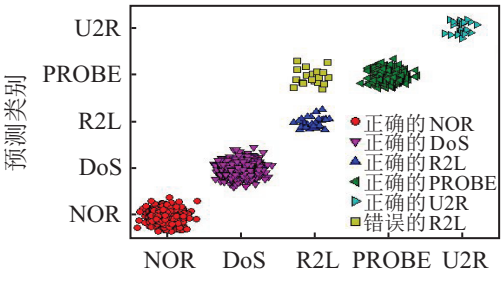
图2 HMGA参数优化进化图

从图2可以看出, 应用HMGA后的种群收敛速度较快, 仅经过10代遗传进化就能达到较为满意的适应度水平, 10代以后的各代平均适应度波动较为稳定, 且最佳适应度一直保持较高水平.

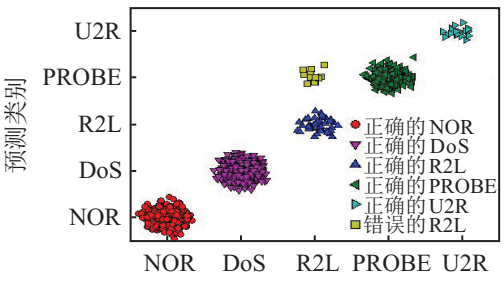
按照1.2节中的算法, HMGA方法得到的最优参数值为: $C=8.283\,1,\sigma=0.085\,926$. 此时5分交叉验证意义下分类准确率为99.920 2%.

4.3 SVM与MBHMSVM分类性能比较

将本文数据集按照4.2节中方法求得的最优参数分别建立单SVM模型和10个SVM基分类器集成的MBHMSVM模型, 并用测试集进行测试. 单SVM和MBHMSVM分类结果的混淆矩阵分别如图3(a)、3(b)所示, 分类结果如表1所示.



(a) SVM分类结果混淆矩阵



(b) MBHMSVM分类结果混淆矩阵

图3 SVM与MBHMSVM分类结果混淆矩阵

表1 SVM与MBHMSVM分类结果对比

方法	错分数	正分数	错分率/%	Kappa统计量
SVM	26	1 969	1.303 3	0.979 2
MBHMSVM	12	1 983	0.601 5	0.990 4

在独立SVM模型26个错分样本中, 原类别全部为R2L(来自远程主机的未授权访问), 被全部错分到类别PROBE(端口监视与扫描). 本文训练集中, R2L为75个样本, PROEBE为380个样本. 可见, 在SVM分类时, 较小样本数据易被视作较大样本数据的噪声数据, 从而形成错分. R2L所代表的故障是众多网络故障问题的主要隐患和来源, 防范意义十分重大, 但分类器错分对故障排除和实施策略产生了较大误导. 10个SVM分类器进行MB集成后, 原本的26个误分样本减少到12个. 可见, 使用MBHMSVM后, 诊断系统故障识别率得到了提升, 网络的安全性得到了增强.

4.4 MBHMSVM分类器性能横向对比

针对MBHMSVM方法与其他主流集成学习方法的性能优劣, 表2给出了不同方法分类误差对比. 使用本文数据集进行实验, 并将所有方法迭代次数均设置为10. 实验结果中, 在相同迭代次数下, MBHMSVM取得了最低的分类错误率.

图4验证了不同迭代次数对分类器的影响, 将MBHMSVM方法、AB-SVM方法、Bagging-SVM方

表 2　不同方法分类误差对比

方法	错分数	错分率/%
MBHMSVM	12	0.601 5
AB-SVM	21	1.052 6
Bagging-SVM	26	1.303 3
RandSubSpace0.5-SVM	26	1.303 3
Rotation Forest-C4.5	26	1.303 3
Rand Forest	26	1.303 3

法在[3, 50]区间内进行整数次迭代. 从图4可以看出, MBHMSVM迭代9或10次时便可达到最小分类误差, 而AB-SVM方法和Bagging-SVM方法在进行10次迭代时的误差率均高于MBHMSVM.

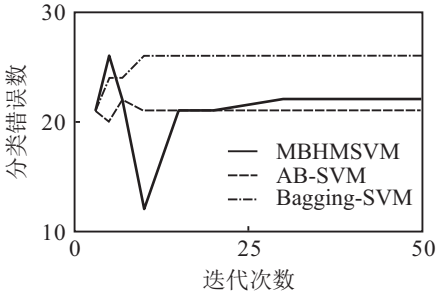


图 4　不同方法迭代次数与错分数间关系

5　结　论

本文基于遗传算法和机器学习方法, 提出了一种MBHMSVM集成学习方法, 遗传算法的优化能力使得分类更加精确, 组合决策分类有效地减小了分类模型偏差. 将该方法应用于网络故障诊断, 可以看出, MBHMSVM方法在KDD99数据集中有着较低的故障识别误差和较高的分类性能, 能够提高网络故障诊断的准确率和可靠性. 同时, MBHMSVM还具有迭代次数少、建模时间短等特点. 在网络故障诊断系统中, 由于监测网络拓扑结构或监测对象实体易发生变化, 经常需要重新建立分类模型, 此时MBHMSVM具有一定使用优势. 目前, 针对复杂系统的故障诊断问题, 还需要进一步地研究如何实现对每个基分类器进行差异性构造以及对大数据的分类处理能力等.

参考文献(References)

[1] 周东华, 胡艳艳. 动态系统的故障诊断技术[J]. 自动化学报, 2009, 35(6): 748-754.
(Zhou D H, Hu Y Y. Fault diagnosis techniques for dynamic systems[J]. Acta Automatica Sinica, 2009, 35(6): 748-754.)

[2] 温祥西, 孟相如, 马志强. 基于双重支持向量机的网络故障诊断[J]. 控制与决策, 2013, 28(4): 506-510.
(Wen X X, Meng X R, Ma Z Q. Network fault diagnosis based on dual-SVM[J]. Control and Decision, 2013, 28(4): 506-510.)

[3] Chen R C, Chen K F. Using rough set and support vector machine for network intrusion detection[J]. Int J of Network Security & Its Application, 2009, 1(1): 1-12.

[4] 唐明珠, 阳春华, 桂卫华. 基于改进的QBC和CS-SVM的故障检测[J]. 控制与决策, 2013, 27(10): 1489-1493.
(Tang M Z, Yang C H, Gui W H. Fault detection based on modified QBC and CS-SVM[J]. Control and Decision, 2013, 27(10): 1489-1493.)

[5] Jayadeva, Khemchandai R. Twin support vector machines for pattern classification[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2007, 29(5): 905-910.

[6] Fung G, Mangasarian O L. Proximal support vector machine classifier[C]. KDD-2001. New York: Association for Computing Machinery, 2001: 77-86.

[7] Freund Y, Schapire R E. A decision theoretic generalization of on-line learning and an application to boosting[J]. J of Computer and System Sciences, 1997, 55(1): 119-139.

[8] Breiman L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123-140.

[9] Bauer E, Kohavi R. An empirical comparison of voting classification algorithms: Bagging, boosting and variants[J]. Machine Learning, 1999, 36(1): 105-139.

[10] Webb G I. Multiboosting: A technique for combining boosting and wagging[J]. Machine Learning, 2000, 40(2): 159-196.

[11] Farhy L S. Modeling of oscillations in endocrine networks with feedback[J]. Methods in Enzymology, 2004, 384: 54-81.

(责任编辑: 齐　霁)