

文章编号:1000-582X(2009)06-674-06

用聚类支持向量机模型监测环保设备运行状态

曹晓莉^{1a,1b},甘思源²,江朝元^{1b}

(1. 重庆工商大学 a. 废油资源化技术与装备教育部工程研究中心;

b. 计算机科学与信息工程学院 重庆 400067;

2. 重庆大学 自动化学院 重庆 400030)

摘要:针对环保设备多状态监测与识别问题,提出了一种基于多分类聚类支持向量机的分类算法模型。该算法模型充分利用多个相关联被监测设备及其状态属性特征,采用神经网络聚类分析获得关联特征的聚类子空间,再对聚类子空间构造二叉树支持向量机进行多分类识别。该算法模型避免了盲目分类,与单纯的支持向量机算法比较具有更好的分类性能。通过对日本大晃船用污水处理装置实测样本的训练、检验表明,该算法具有 98% 以上的分类识别率。

关键词:支持向量机;设备状态;多分类;聚类

中图分类号:TP181

文献标志码:A

Running condition monitoring for environmental protection equipment based on a clustering support vector machine

CAO Xiao-li^{1a,1b}, GAN Si-yuan², JIANG Chao-yuan^{1b}

(1a. Engineering Research Center for Waste Oil Recovery Technology and Equipment,

b. College of Computer Science and Information Engineering, Chongqing

Technology and Business University, Chongqing 400067, P. R. China;

2. College of Automation, Chongqing University, Chongqing 400030, P. R. China)

Abstract: A multi-classification clustering model based on Support Vector Machines (SVM) is proposed to address problems of condition monitoring and identification in environmental protection equipment. Based on the associated attributes of monitoring equipment and their condition features, a neural network is adopted for clustering analysis to obtain clustering subspaces. The clustering subspace is employed to design a binary tree architecture based multi-classification SVM for multi-classification recognition. Compared with the traditional SVM learning algorithms, the proposed algorithm improves the classification performance while avoiding blind classification. The model is applied to train and examine the measured samples for a ship sewage treatment device made in Japan. The results show that the classified recognition rate of this algorithm exceeds 98% and is valuable for practical applications.

Key words: support vector machines; equipment condition; multi-classification; cluster

收稿日期:2009-02-12

基金项目:重庆市教委资助项目(KJ080708);废油资源化技术与装备教育部工程研究中心重大项目(07011301)

作者简介:曹晓莉(1970-),女,重庆工商大学副教授,主要从事模式识别与智能控制、智能传感器方向研究,

(Tel)13637879839;(E-mail)cxlsml@163.com.

欢迎访问重庆大学期刊社网 <http://qks.cqu.edu.cn>

环保处理设备是用于污染处理的环保装置,其运行状态直接关系到对污染源处理的有效性,是环保监管与行政执法的重要依据。准确、有效的监测环保处理设备的运行状态具有十分重要的现实意义。在实际工程应用中,由于环保处理设备要消耗一定的能耗,经常出现人为停机、空载运行或检查时启动等违规现象,环保处理设备未发挥其应有的作用。为此,有效监测环保处理设备的运行状态是保证处理设备正常运行的重要手段。对环保处理设备的监测不仅需要监测环保处理设备自身的运行状态,同时也需要监测生产设备的运行状态,并在时域空间对关联设备的多个参数进行监测、融合与识别,才能准确判定目标监测设备的真实运行状态。此外,由于人为作弊、传感器故障、设备自身故障等情况可能出现,使设备状态监测问题更加复杂化,需要在一定的信息冗余基础上,获得关联设备之间以及设备自身状态之间在时域空间内的状态特征才能识别设备的停止、运行、故障和作弊等状态。

针对设备运行状态识别分类问题,不少学者采用小波分析^[1]、遗传算法^[2]、神经网络^[3]和支持向量机(support vector machines, SVM)^[4-6]等人工智能学习方法取得了相应的研究成果。但对更加复杂性的多个相关设备运行状态监测与样本有限的分类问题,上述人工智能算法则不能很好解决。为此,结合支持向量机算法具有对有限样本的信息进行分类求最优解,在理论上局部最优解就是全局最优解以及可构造核函数将非线性样本空间转换到高维线性空间,以在高维线性空间构造线性分类函数来实现对原样本空间的非线性分类等优点^[7-9],提出了一种基于多分类聚类支持向量机的设备运行状态监测方法。

1 设备运行状态及其相关性

如图 1 所示,船用污水处理装置和船舶都是系统监测对象,其中船用污水处理装置是目标监测对象。监测系统由现场监测终端和监测系统构成,其中监测终端部署在船舶上,通过若干传感器监测目标对象设备的电压、电流、风压和液位等参数;同时也监测船舶上 GPS 定位仪输出的船舶航行速度和监测船舶的轮机供电情况。监测终端将各监测传感器参数上传给系统,由系统实现在时域空间的设备状态监测。

设备与状态的相关性约束为:

1)若船舶航行,则必然产生生活污水,且要求处理设备必须启动运行,否则可暂停。

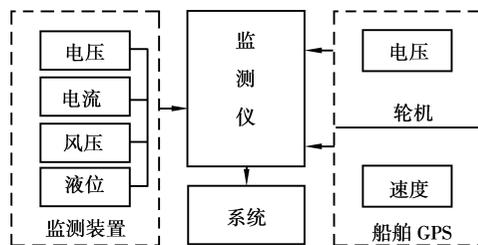


图 1 SOFM-SVM 分类模型设计

2)若船舶停泊,节电停机时间不得超过 T_s/h , 否则导致内部微生物死亡。

3)若设备运行,则必须向送风供氧,以保持内部微生物存活,且风压力不低于 p_d , 否则运行无效或传感器故障。

4)若设备运行,则污水水位应周期性 T_L 变化, 否则可能作弊运行或传感器故障。

5)若装置运行,则工作电流不低于 a_d , 否则可能空载运行或传感器故障。

6)若轮机供电而设备停电,则可能停机或设备故障。

2 SOFM-SVM 分类模型设计

设 u_d 、 a_d 、 p_d 、 l_d 分别为监测设备的电压、电流、压力和液位, u_s 、 v_s 分别为船舶轮机电压和航行速度;识别设备状态为停机、运行、故障和作弊 4 种状态。该问题可抽象为属性集 X 到状态分类集 Y 的映射,即参数属性 $X = \{u_d, a_d, p_d, l_d, u_s, v_s\}$, 状态 $Y = \{1, 2, 3, 4\}$ 的多分类支持向量机分类寻优问题。

考虑到传感器故障与人为作弊行为的相似性和与正常状态之间的差异性,以及分类目标必须避免将非违规状态分类为违规状态的要求。若直接构造多个 SVM 进行状态分类,将难以避免对非违规状态的误分类,同时也无法充分发挥 SVM 较强的推广能力。

笔者利用多设备间相关性参数的信息冗余性构造了如图 2 所示的聚类支持向量机分类模型。该模型首先采用自组织特征映射算法(self-organizing feature maps, SOFM)进行神经网络聚类分析,将样本间差异性较大的样本聚类,输出若干样本子集^[10-11];再将各样本子集由核函数映射到高维特征空间构造多个 SVMs,将 SVMs 转化为二叉树 SVMs 实现多类问题分类,并根据不同样本子集选取二叉树 SVMs 的边际系数^[12]。此算法模型简称 SOFM-SVM 模型。

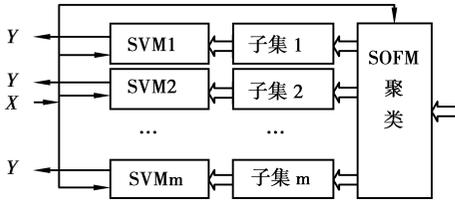


图 2 SOFM-SVM 分类模型

3 基于 SOFM-SVM 的监测状态分类

3.1 SOFM 聚类分析

SOFM 是一种神经网络聚类分析方法,由输入层和竞争层组成,可在无监督方式进行网络训练,通过网络结构的自组织从输入样本集中发现并提取其内在的特征,在网络输出结点权向量空间上形成输入分布拓扑图,反映输入数据的分布规律,自动对输入样本进行聚类。

设 SOFM 神经网络输入层模式向量为 $\mathbf{X}_i = (x_{i1}, x_{i2}, \dots, x_{im})$, 输入层神经元数为 $n, i=1, 2, \dots, q$; 竞争层向量为 $\mathbf{A}_j = (a_{j1}, a_{j2}, \dots, a_{jm})$, 竞争层神经元数为 $m, j=1, 2, \dots, m$; 竞争神经元与输入层神经元之间的连接权向量为 $\mathbf{W}_j = (w_{j1}, w_{j2}, \dots, w_{jn}), j=1, 2, \dots, m$ 。

SOFM 聚类算法为

1) 初始化: 网络连接权向量 $\mathbf{W}_{ij} (i=1, 2, \dots, n; j=1, 2, \dots, m)$ 为区间 $[0, 1]$ 的随机值; $t=0; i=0; j=0$; 初始学习率取 $0 < \eta(0) < 1$ 随机值。

2) 输入模式向量 $\mathbf{X}_i = (x_{i1}, x_{i2}, \dots, x_{im}), i=1, 2, \dots, k$; 并对 \mathbf{X}_i 和 \mathbf{W}_j 进行归一化为 $\overline{\mathbf{x}}_i$ 和 $\overline{\mathbf{w}}_{ij}$;

3) 计算输入与所有输出结点连接向量 \mathbf{w}_j 的欧氏距离

$$d = \left[\sum_{i=1}^n (\overline{\mathbf{x}}_{ik} - \overline{\mathbf{w}}_{ij})^2 \right]^{\frac{1}{2}}, \quad j = 1, 2, \dots, m. \quad (1)$$

4) 找出最小距离 d_k , 即为竞争获胜结点 g

$$d_k = \min(d_j), \quad j = 1, 2, \dots, m. \quad (2)$$

5) 对竞争层获胜输出结点 A_g 的几何邻域 $N_g(t)$ 内所有结点连接的权值进行修正调整

$$\overline{\mathbf{w}}_{ij}(t+1) = \overline{\mathbf{w}}_{ij}(t) + \eta(t) [\overline{\mathbf{x}}_{ik} - \overline{\mathbf{w}}_{ij}(t)], \quad (3)$$

$$j \in N_g(t), i = 1, 2, \dots, n, j = 1, 2, \dots, m,$$

$$0 < \eta(t) < 1,$$

令学习模式 $i=i+1$, 返回步骤 3), 直到模式学习结束。

7) 更新学习率 $\eta(t)$ 和邻域 $N_g(t)$

$$\eta(t) = \eta(0) \left(1 - \frac{t}{T} \right), \quad (4)$$

其中 $\eta(0)$ 为初始学习率; t 为学习次数; T 为预定总学习次数。

$$N_g(t) = \text{INT} \left[N_g(0) \left(1 - \frac{t}{T} \right) \right]. \quad (5)$$

其中 $N_g(0)$ 为初始邻域。

8) 令学习次数 $t=t+1$, 返回步骤 2), 直至 $t=T$ 为止。

算法中, 只要调整邻域 $N_g(t)$ 内的连接向量权值, 使权向量 \mathbf{w}_{ij} 更接近或更偏离输入模式向量 \mathbf{X}_i , 这一调整过程就是竞争学习过程。随着不断的学习, 所有连接权向量都在输入向量空间相互分离, 形成各自代表输入向量空间的一类模式, 这就是 SOFM 的自动聚类分析。

3.2 二叉树 SVM 状态分类

经典 SVM 算法是针对二类分类问题, 对于多分类问题, Weston 等人在 1998 年提出了基于经典二分类 SVM 重构多值分类模型, 但这类算法的目标函数十分复杂, 实现比较困难^[13]。更多的学者是将多类分类逐步转化为二类分类, 用多个二分类 SVM 组成多类分类器, 其中 Vapnik 提出了 1 对多 (one-against-rest) 算法, Kressel 提出了 1 对 1 (one-against-one) 算法, 此两种算法可以将复杂的多分类问题简化处理, 避免了目标函数的复杂化, 但这两种算法均存在计算量较大、推广误差和不可分等问题^[14-16]。为此, 更为广泛应用的是二叉树 SVM, 它是基于 1 对多 SVM 算法和改进, 针对 N 分类问题需要 $N-1$ 个二分类 SVM, 且每个输入样本不一定需要遍历全部二分类 SVM 计算, 也不存在不可分的问题。

针对研究的问题, 设分类训练样本集为 $(\mathbf{x}_i, \mathbf{y}_i), i=1, 2, \dots, n, \mathbf{x}_i \in R^d, \mathbf{y}_i \in \{1, 2, 3, 4\}$, 式中 n 为训练样本个数; $d=6$, 为每个样本向量的维数; $\mathbf{y} = \{1, 2, 3, 4\}$ 为目标分类类别符号标量。此问题为 4 分类问题, 为提高分类速度和避免不可分类, 采用二叉树 SVM 算法实现多类状态分类, 模型如图 3 所示。

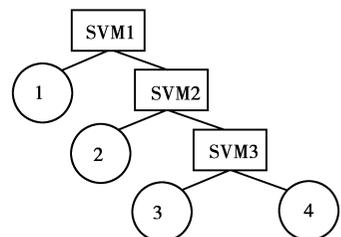


图 3 二叉树支持向量机模型

3.2.1 二叉树 SVM 模型

设备运行状态分类过程为:将输入设备状态参数向量 $\mathbf{x}_i(t)$ 到根节点 SVM1 分类器,状态 1 被分类出来时 $y_i=1$ 分类停止,未分类出来 $y_i=-1$,继续输入到 SVM2 分类器;若状态被分类出来时 $y_i=2$ 分类停止,否则 $y_i=-1$,继续输入到 SVM3 分类器;若被 SVM3 分类为状态 3 则 $y_i=3$,否则 $y_i=4$ 。由此可见:4 分类设备状态问题仅需要 3 个二分类 SVM,且分类过程在被分类出来时即中止计算,具有很好的计算效率。

基于 SOFM-SVM 模型进行状态分类,需要构造 4 个二叉树 SVM 分类器,分别针对 4 类状态聚类子空间进行分类,每个二叉树 SVM 分类器通过机器学习由不同的边际系数来确定。

3.2.2 二叉树 SVM 算法

由 SVM 理论可知,经典二分类 SVM 就是在向量空间寻找一个最优分函数并满足:

$$y_i(\mathbf{w} \cdot \mathbf{x} + b) \geq 1; i = 1, 2, \dots, m, \quad (6)$$

其中: $\mathbf{w} \in R^n$ 为权值向量; b 为偏置标量系数。

通过引入拉格朗日乘子 α_i 在对偶空间求解得到最优分类函数

$$f(x) = \text{sign}\left(\sum_{i=1}^{SV} \alpha_i y_i x^T \cdot \mathbf{x} + b\right), \quad (7)$$

式中: SV 为所有支持向量(即 $\alpha_i > 0$ 所对应的样本)的集合。

对于非线性分类问题, SVM 可将一个非线性样本 x_i 映射成线性样本 $\Phi(x)$, 原空间 R^n 映射到一个高维的特征空间 Ω , 再在高维特征空间 Ω 求最优分类函数

$$f(x) = \text{sign}\left[\sum_{i=1}^{SV} \alpha_i y_i \Phi(x) \cdot \Phi(x) + b\right], \quad (8)$$

SVM 可利用满足 Mercer 条件的核函数 $K(x, \mathbf{x}')$ 来替代式(8)中的 $\Phi(x) \cdot \Phi(x)$ 点积, 此时最优分类函数为

$$f(x) = \text{sign}\left[\sum_{i=1}^{SV} \alpha_i y_i K(x, \mathbf{x}') + b\right]. \quad (9)$$

核函数的引入使得操作可以直接在输入空间进行而不必在潜在的高维特征空间进行, 这种方式可以避免维数灾难。常用的核函数有: 多项式核函数、径向基核函数(radial basis function)、多层核函数、Fourier 核函数、附加核函数和张量积核函数等^[18], 其中对于小样本的机器学习径向基核函数具有突出优势

$$K(x, \mathbf{x}') = \exp(-\gamma \|x - \mathbf{x}'\|^2), \quad (10)$$

式中 γ 是一个经验可调的控制半径。

根据二叉树 SVM 模型, 在要建立的 K 个分类器中的 m 个 SVM 分类函数为

$$f^m(x) = \text{sign}\left[\sum_{i=1}^{SV} \alpha_i^m y_i^m K(x, \mathbf{x}') + b^m\right]. \quad (11)$$

3.2.3 二叉树 SVM 训练

由图 3 所示的二叉树 SVM 模型可知, 研究的问题需要通过样本训练获得 3 个二分类 SVM, 由图 2 所示分类模式可知需要训练 4 个二叉树 SVM, 共 12 个二分类 SVM。具体训练步骤为:

- 1) 根据实际系统要求实测采集参数向量 \mathbf{x}_i 及确定训练样本个数; $k=1, T=1$;
- 2) 选择核函数为 RBF, 固定 γ 系数和边际系数 C ;
- 3) 计算式(11)的分类函数中 α_i, b

$$Q(a) = \sum_i^m a_i - \frac{1}{2} \sum_{i,j=1}^m a_i a_j y_i y_j e^{-\gamma \|x - x'\|^2}, \quad (12)$$

约束条件为

$$\sum_{i=1}^m y_i a_i = 0; C \geq a_i \geq 0; i = 1, 2, \dots, m,$$

$$b = \frac{1}{|U|} \sum_{i \in U} (y_i - \mathbf{w} x_i). \quad (13)$$

- 4) 对 RBF 核函数不断修改其参数 γ , 对式(12)的计算不断修改 C , 用训练样本进行训练学习, 以获得在固定值下结果最佳的核函数参数 γ 和 C 。

- 5) 在最佳核函数情况下确定参数 α_i 和 b 作为核函数参数;

- 6) $k=k+1$, 再返回 2) 训练二叉树 SVM 模型中第 m 个 SVM, 直到 $k=3$;

- 7) $T=T+1$, 再返回 2) 训练第 T 个二叉树 SVM 模型, 直到 $T=4$ 。

4 SOFM-SVM 模型检验

训练样本和检验样本数据均来自日本大晃的船用污水处理装置现场实测数据。现场状态监测仪以 10 s 为周期自动采集停泊、航行、停电、供电 4 种关联设备状态下并包含了监测目标设备的停机、运行、故障和作弊 4 种状态的样本数共 1 440 个, 其中训练样本数为 440 个, 检验样本为 1 000 个。

1) SOFM 分析检验

经 SOFM 聚类分析器分析, 将 1 000 个样本构成设备状态属性集 X 分为了 4 个状态属性子集: $X_{\text{停泊供电}}$ 、 $X_{\text{航行供电}}$ 、 $X_{\text{航行停电}}$ 、 $X_{\text{停泊停电}}$, 聚类分析结果如表 1 所示。

表1 SOFM 检验输出结果

u_d/v	a_d/Λ	p_d/KPa	l_d/m	u_s/V	$v_s/(km \cdot h^{-1})$	环境状态	样本数
380~398	8.0~12.8	0.0~25.7	1.0~2.1	380~398	0.0~0.3	停泊供电	250
380~398	8.0~12.8	0.0~25.7	1.0~2.1	380~398	0.4~26.3	航行供电	250
0	0	0	0	380~398	22.4~26.3	航行停电	250
0	0	0	0	380~398	0.0~0.3	停泊停电	250

2) SVM 分类检验

为检验分类方法的有效性,对 1 000 个检验样本分别采用二叉树 SVM 和 SOFM-SVM2 种算法进行检验,表 2 给出了不同的检验结果。

表2 SVM 与 SOFM-SVM 的检验结果

运行状态	SVM		SOFM-SVM	
	时间/ms	分类率/%	时间/ms	分类率/%
S1=作弊	127	78.6	174	98.4
S2=故障	261	85.8	388	99.7
S3=停机	389	100.0	503	100.0
S4=运行	503	64.4	676	98.1

由对比检验结果可见

①SVM 算法采用二叉树模型,计算越多分类时间越长;

②单纯的 SVM 算法没有充分利用设备相关属性参数,S1、S2 状态被误分到 S4,除 S3 外其它状态误分类率较高;

③SOFM-SVM 算法充分利用了设备相关属性参数,仍存在将 S1、S2 误分到 S4 的情况,但误分类率显著下降,达到可实用化水平;

④SOFM-SVM 算法由于模型更加复杂,分类时间要明显大于单纯的 SVM 方法。

5 结 语

1)从上述样本分析验证可知,SOFM-SVM 算法充分利用了监测设备及其关联设备状态属性特征,减少一级 SVM 分类器,降低了计算量,一定程度上避免了分类的盲目性;

2)从算法模型检验结果可知,与纯粹的 SVM 分类法相比较,SOFM-SVM 算法模型具有良好的

分类正确率和较好的泛化性能;

3)针对不同的聚类子空间,交叉检验各种核函数,是一个值得研究的问题,以进一步提高分类模型的推广能力。

参考文献:

- [1] 李淑娥,邸敏艳,吕锋. 基于小波变换的电气设备故障诊断技术[J]. 仪表技术与传感器,2004(5):46-48.
LI SHU-E, DI MIN-YAN, LU FENG. Technology of fault diagnosis in electricity equipments based on wavelet transformation[J]. Instrument Technology and Sensor,2004(5):46-48.
- [2] 庄哲民,肖广辉,曹勤. 基于遗传神经网络的异步电动机故障诊断研究[J]. 测试技术学报,2004,18(4):377-380.
ZHUANG ZHE-MIN, XIAO GUANG-HUI, CAO QIN. Fault detection of asynchronism motor based on genetic neural networks [J]. Journal of Test and Measurement Technology, 2004,18(4):377-380.
- [3] 庄哲民,林志强. 基于神经网络的滚动轴承故障诊断研究[J]. 仪器仪表学报,2000,21(1):73-82.
ZHUANG ZHE-MIN, LIN ZHI-QIANG. Research on ball bearing failure diagnosis based on neural network [J]. Chinese Journal of Scientific Instrument, 2000, 21(1):73-82.
- [4] 王红军,徐小力. 支持向量机在设备故障诊断方面的应用研究概述[J]. 机械设计与制造,2005(9):157-159.
WANG HONG-JUN, XU XIAO-LI. Overview of the support vector machines and its application research on equipment fault diagnosis [J]. Machinery Design & Manufacture[J]. 2005(9):157-159.
- [5] JIANG Z Q, FU H G, LI L J. Support vector machine for mechanical faults classification [J]. Journal of Zhejiang University Science, 2005,6A(5): 433-439.
- [6] CRISTIANINI N, TAYLOR J S. An introduction to

- support vector machines and other kernel based-learning methods [M]. Beijing: Publishing House of Electronics Industry, 2004.
- [7] CAI Y Z, HU Z H, YIN R P, et al. Combination of multi-class probability support vector machines for fault diagnosis [J]. Journal of Donghua University, 2006, 23 (1):12-17.
- [8] 张周锁,李凌均,何正嘉. 基于支持向量机的多故障分类器及应用[J]. 机械科学与技术, 2004, 23 (5): 536-601.
- ZHANG ZHOU-SUO, LI LING-JUN, HE ZHENG-JIA. Multi-fault classifier based on support vector machine and its applications[J]. Mechanical Science and Technology, 2004, 23(5):536-601.
- [9] 李东晖,杜树新,吴铁军. 监控污水处理过程运行状态的加权支持向量机方法[J]. 化工自动化及仪表, 2005, 32 (1): 18-20.
- LI DONG-HUI, DU SHU-XIN, WU TIE-JUN. Weighted support vector machine for operation monitoring of wastewater treatment processes [J]. Control and Instruments in Chemical Industry, 2005, 32 (1): 18-20.
- [10] 林金山,林敏. 神经网络的特征映射聚类算法研究[J]. 现代电子技术, 2006(24):41-43.
- LIN JIN-SHAN, LIN MIN. Research in clustering of SOFM neural network [J]. Modern Electronics Technique, 2006(24):41-43.
- [11] HU Z H, CAI Y Z, LI Y G, et al. Data fusion for fault diagnosis using multi-class support vector machines [J]. Journal of Zhejiang University Science, 2005, 6(10):1030-1039.
- [12] HSU C W, LIN C J. A comparison of methods for multi-class support vector machines [J]. IEEE Transactions on Neural Networks, 2002, 13 (3): 415-425.
- [13] VAPNIK V N. Statistical learning theory [M]. New York: John Wiley & Sons Inc, 1998.
- [14] ZHENG S B, TANG H J, HAN Z Z, et al. Solving large-scale multiclass learning problems via an efficient support vector classifier [J]. Journal of Systems Engineering and Electronics, 2006, 17(4):910-915.
- [15] ZANGHIRATI G, ZANNI L. A parallel solver for large quadratic programs in training support vector machines [J]. Parallel Computing, 2003, 29 (4): 535-551.
- [16] 王红军,徐小力. 支持向量机理论在设备状态趋势预测上的应用研究[J]. 兰州理工大学学报, 2005, 31(6): 36-39.
- WANG HONG-JUN, XU XIAO-LI. Applicational study of support vector machine theory for machinery condition prediction[J]. Journal of Lanzhou University of Technology, 2005, 31(6):36-39.
- [17] 陶秀凤,周鸣争. 基于支持向量机的多传感器信息融合算法[J]. 计算机技术与发展, 2006, 16(6):177-183.
- TAO XIU-FENG, ZHOU MING-ZHENG. An algorithm of multiple sensor Information fusion based on SVM[J]. Computer Technology and Development, 2006, 16(6):177-183.
- [18] CRISTIANINI N, TAYLOR J S. An introduction to support vector machines and other kernel-based learning methods [M]. Cambridge: Cambridge University Press, 2000.

(编辑 侯湘)