第 24 卷第 1 期
2007 年 02 月

控 制 理 论 与 应 用
Control Theory & Applications

Vol. 24 No. 1
Feb. 2007

文章编号：1000−8152(2007)01−0064−05

基于聚类和支持向量机的非线性时间序列故障预报

张军峰, 胡寿松

(南京航空航天大学 自动化学院, 江苏 南京 210016)

摘要: 针对非线性时间序列故障预报问题, 提出了一种基于聚类和支持向量机的方法. 将正常的时间序列按照K-均值聚类算法进行聚类学习, 同时利用支持向量机回归的时间序列预测算法获得预测序列, 然后通过比较聚类所得的正常原型和预测序列的相似性实现故障预报. 仿真结果表明: 本文提出的方法更能满足实时性的要求, 也更为准确.

关键词: 故障预报; K-均值聚类; 支持向量回归; 时间序列预测

中图分类号: TP273　　**文献标识码**: A

Nonlinear time series fault prediction based on clustering and support vector machines

ZHANG Jun-feng, HU Shou-song

(College of Automation Engineering, Nanjing University of Aeronautic and Astronautic, Nanjing Jiangsu 210016, China)

Abstract: Based on clustering and support vector machines, a new method is proposed to solve the nonlinear time series fault prediction. The normal time series is clustered using K-means clustering algorithm to get the normal prototype. Meanwhile, the predicting series is obtained by time series predicting algorithm based on support vector regression. Fault prediction can also be implemented by calculating the similarity between the normal prototype and the predicting series. Finally, the simulation results indicate that the proposed method can predict the fault more quickly and more accurately.

Key words: fault prediction; K-means clustering; support vector regression; time series prediction

1　引言(Introduction)

随着科技不断进步, 现代化的工程系统日趋复杂. 一些对可靠性要求很高的系统(如航天航空系统), 不仅要求能够在出现故障时进行准确地诊断, 还希望对系统劣化趋势作出早期预报,以便减少故障所造成的破坏, 这就是故障预报. 它是增强诊断系统的故障早期发现能力, 提高实时性的一种重要手段.

到目前为止, 虽然国内外对故障预报技术的研究成果还不是很多, 但是还是有许多学者做了很多开创性的工作. Ho和Xie[1]针对故障发生的时间间隔构成的数据序列建立ARIMA模型, 采用经典时间序列分析的方法, 建立预报公式, 对下一次故障发生的时刻作出预报. 但这种方法用线性模型来拟合数据序列, 故从本质上来看,它并不适合预报非线性系统. Tse和Atherton[2]采用了回归神经网络对香港一家化工厂的冷却塔的鼓风机进行故障预报. 虽然运用神经网络进行故障预报取得了一定的效果, 但是神经网络方法是一种“黑盒”方法, 无法表达和分析系统的输入与输出间的关系. 而且没有通用的准则确定网络的结构.

一个完整的故障预报系统应满足如下设计要求: 实时性、正确性和完备性. 而目前众多的故障预报方法大都通过求取预测值与实际观测值的差值, 然后用该差值与设定的阈值相比较, 得出判断. 显然, 这样做很难满足实时性的要求. 本文提出了一种新的故障预报方法, 即融合聚类技术和支持向量机回归预测技术的故障预报方法, 可以解决上述问题.

2　故障预报系统 (Fault prediction system)

基于聚类和支持向量机回归预测的故障预报系统主要有3个模块组成: 正常原型提取模块、回归预测模块和相似性度量模块, 其系统的结构图如图1所示.

由于本文根据时间序列进行故障预报, 因此图1中正常原型提取模块主要是对正常时间序列运用聚类算法, 得出正常原型使其能够反映全部正

收稿日期: 2005−08−01; 收修改稿日期: 2006−02−23.
基金项目: 国家自然科学基金重点资助项目(60234010); 航空科学基金资助项目(05E52031); 国防基础科研资助项目(K1603060318).

常时间序列的特征; 回归预测模块是利用基于时间序列回归的算法建立预测模型, 然后进行时间序列的预测; 而相似性度量模块则是依据时间序列的相似性, 预测故障是否会发生.

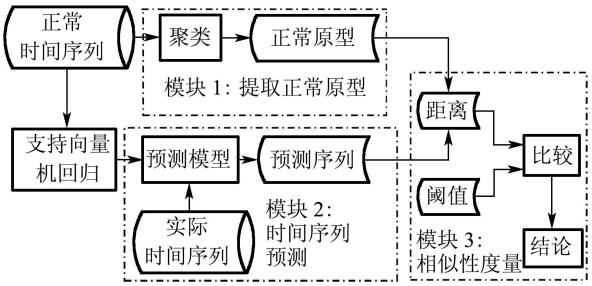


图 1　故障预报系统

Fig. 1　Fault prediction system

2.1　原型提取 (Prototype extraction)

聚类是一个非常重要的非监督学习问题, 它可以简单地定义为: 将目标分组, 使得组内成员之间存在某种相似性的过程. 本文的故障预报系统加入基于聚类算法的正常原型提取模块主要由于目前故障预报方法大都很难满足实时性的要求. 实际系统在正常运行时, 其状态信息一般是呈现周期性变化或者在一定范围内变化, 因此可以在考察正常状态信息规律的基础上, 提取其原型, 用以和预测值进行比较判断.

本文运用K-均值算法[3]对系统运行的正常状态信息进行聚类. K-均值聚类是最简单的非监督学习算法, 它基于误差平方和准则[4].

2.2　时间序列预测 (Time series prediction)

由Vapnik[5]创立的支持向量机(SVM)在机器学习中表现出了十分独特的特征, 它首先应用于模式分类中, 然后推广到了回归和时间序列预测[6]等领域. 与常规神经网络法所基于的经验风险最小化原理不同, SVM遵循了结构风险最小化原理, 这就使得SVM具有更好的泛化性能. 由于集成了最大间隔超平面、 Mercer核、凸二次规划、稀疏解和松弛变量等多项技术, SVM方法可以求得全局最优解.

2.2.1　支持向量机回归(Support vector regression)

设给定的训练样本为

$$\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_l, y_l)\} \subset \mathbb{R}^n \times \mathbb{R}.$$

首先用一个非线性映射ϕ 把数据映射到一个高维特征空间, 然后在高维特征空间中进行线性回归. 设回归函数为

$$f(\boldsymbol{x}) = \langle \boldsymbol{w}, \phi(\boldsymbol{x}) \rangle + b. \tag{1}$$

选取ε 不敏感损失函数, 根据结构风险最小化原则, 引入松弛变量, 回归问题可归结为在不等式约束下最小化函数:

$$\frac{1}{2} \parallel \boldsymbol{w} \parallel^2 + C \sum_{i=1}^{l} (\xi_i + \xi_i^*). \tag{2}$$

应用拉格朗日乘子法求解这个具有线性不等式约束的二次规划问题, 可得到该优化问题的对偶形式. 此时的决策函数具有如下形式:

$$f(\boldsymbol{x}) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) K(\boldsymbol{x}_i, \boldsymbol{x}) + b. \tag{3}$$

其中$K(x_i, x_j)$ 被称作核函数. 核函数的引入, 可以避免在高维特征空间中进行复杂的运算. 而且任何函数只要满足Mercer 条件, 就可以作为核函数[7]. 常用的核函数一般包括多项式核和径向基核.

2.2.2　超参数选择 (Hyper-parameter selection)

支持向量机的泛化性能取决于超参数C, ε以及核参数的选择. C 取得小, 则对样本数据中超出ε管道的样本惩罚就小, 使训练误差变大. C 取得大, 系统的泛化能力变差. 同样, ε选小, 回归估计精度高, 但支持向量数量增多, ε选大, 回归估计精度降低, 支持向量数量少, SVM 的稀疏性大. 在实际问题求解时, 如何合适地选取这些参数, 目前没有有效的方法, 一般都是将SVM的超参数预先设定, 这就避免不了随意性. 也有用交叉验证技术进行参数选择, 这样无疑增加了额外的计算开销. 在本文中, 将直接考察训练样本, 并根据其样本特性选择参数.

根据文献[8], 将惩罚系数C确定为训练数据的输出范围. 为了防止其对奇异点的过分敏感, C值可以确定为如下的形式: $C = \max(|\bar{y} + 3\sigma_y|, |\bar{y} - 3\sigma_y|)$, 其中$\bar{y}$ 表示输出的平均值, σ_y表示标准差. 而ε 与输入的噪声是成比例的[9], 即: $\varepsilon \sim \sigma_{\text{noise}}/\sqrt{n}$. 本文选取径向基核作为核函数,则核参数即为径向基核的扩展参数. 由于扩展参数一般反映了输入样本的范围, 因此可以将扩展参数选为如下的形式: $\sigma \sim (0.1 \sim 0.5) \cdot \text{range}(\boldsymbol{x})$.

2.2.3　基于支持向量机回归的时间序列预测 (Time series prediction based on SVR)

使用支持向量机进行时间序列预测, 其关键问题在于如何重构线性空间, 找到输入输出的对应关系. 如果x_{t+p} 是预测的目标值, 将先前的目标值$\{x_t, x_{t+1}, \cdots, x_{t+p-1}\}$ 作为输入值, 建立预测模型, 其中$t = 1, \cdots, n-p$. 经过变换, 可以得到用于支持向量机学习的样本:

$$X = \begin{bmatrix} x_1 & x_2 & \cdots & x_p \\ x_2 & x_3 & \cdots & x_{p+1} \\ \vdots & \vdots & & \vdots \\ x_{n-p} & x_{n-p+1} & \cdots & x_{n-1} \end{bmatrix}, \; Y = \begin{bmatrix} x_{p+1} \\ x_{p+2} \\ \vdots \\ x_n \end{bmatrix}. \tag{4}$$

可以看出, 输入与输出之间存在着一一映射关系, 其中$f:\mathbb{R}^p\to\mathbb{R}$, 其中$p$值称为嵌入维数. 本文采用最终预报误差(FPE)准则评价预测误差.

$$\mathrm{FPE}(k)=\frac{n+k}{n-k}\sigma_{\mathrm{e}}^2=\frac{n+k}{(n-k)^2}\sum_{t=k+1}^{n}(x_t-\hat{x}_t)^2,$$

其中$x_t,\hat{x}_t$分别表示实际值和预测值. 当k变化时, $\mathrm{FPE}(k)$ 也相应发生变化, 因此可以找到一个最优值k_{opt}使得FPE达到最小, 此时嵌入维数$p=k_{\mathrm{opt}}$.

在得到输入输出样本和嵌入维数之后, 就可对支持向量机进行训练, 得到的回归函数表示:

$$y_t=\sum_{i=1}^{n-p}(\alpha_i-\alpha_i^*)\cdot K(\boldsymbol{x}_i\cdot\boldsymbol{x}_t)+b. \tag{5}$$

这样就可以得到对第$n+1$点的预测值:

$$y_{n+1}=\sum_{i=1}^{n-p}(\alpha_i-\alpha_i^*)\cdot K(\boldsymbol{x}_i\cdot\boldsymbol{x}_{n-p+1})+b. \tag{6}$$

如此可以得到一个样本数据:

$$\boldsymbol{x}_{n-p+2}=\{x_{n-p+2},x_{n-p+3},\cdots,x_n,\hat{x}_{n+1}\},$$

其中$\hat{x}_{n+1}$表示第$n+1$个数据的预测值, 即$\hat{x}_{n+1}=y_{n+1}$.

同样, 可以得到第$n+2$点的预测值:

$$y_{n+2}=\sum_{i=1}^{n-p}(\alpha_i-\alpha_i^*)\cdot K(\boldsymbol{x}_i\cdot\boldsymbol{x}_{n-p+2})+b. \tag{7}$$

一般地,第l步的支持向量机预测模型为

$$y_{n+l}=\sum_{i=1}^{n-p}(\alpha_i-\alpha_i^*)\cdot K(\boldsymbol{x}_i\cdot\boldsymbol{x}_{n-p+l})+b. \tag{8}$$

其中$\boldsymbol{x}_{n-p+l}=\{x_{n-p+l},\cdots,x_{n+1},\cdots,x_{n+l-1}\}$.

2.3 相似性度量 (Similarity measure)

本文提出的故障预报方法, 通过K-均值聚类提取正常原型, 根据支持向量机回归的时间序列预测算法得到预测序列, 然后考察正常原型和预测序列的相似性进行故障预报.

最简单和最常用的时间序列的相似性度量是计算两个时间序列之间的Euclidean距离. 预测序列$\boldsymbol{x}_P$与第i个正常原型$\boldsymbol{x}_{Fi}$之间的Euclidean距离定义为:

$$D_{\mathrm{E}}(\boldsymbol{x}_P,\boldsymbol{x}_{Fi})=\sqrt{(\boldsymbol{x}_P-\boldsymbol{x}_{Fi})\cdot(\boldsymbol{x}_P-\boldsymbol{x}_{Fi})^{\mathrm{T}}}. \tag{9}$$

如果样本集拥有紧凑的或者孤立的聚类中心时, 采用Euclidean距离来度量时间序列的相似性能够取得很好的效果. 但是直接使用Euclidean距离的弊端在于具有较大数值的特征会占据主导地位, 而且各特征之间的相关性会影响相似性的度量[3], 此时可以选择Mahalanobis距离.通常Mahalanobis距离是两个或多个相关变量所定义空间中两点间的距离. 预测序列$\boldsymbol{x}_P$与第i个正常原型$\boldsymbol{x}_{Fi}$之间的Mahalanobis距离定义为:

$$D_{\mathrm{M}}(\boldsymbol{x}_P,\boldsymbol{x}_{Fi})=\sqrt{(\boldsymbol{x}_P-\boldsymbol{x}_{Fi})\mathrm{Cov}^{-1}(\boldsymbol{x}_P-\boldsymbol{x}_{Fi})^{\mathrm{T}}}. \tag{10}$$

其中Cov表示k个正常原型的协方差矩阵.

当预测序列和聚类序列之间的Euclidean距离或Mahalanobis距离超过一定的阈值时, 就可以认定即将发生故障,从而进行了故障预报.

3 仿真结果 (Simulation results)

由于实际系统运行时, 其正常的状态信息一般都呈现周期性变化或者在一定范围内变化, 因此本文选取Henon 时间序列与釜式反应器作为故障预报的实例. 产生的两个时间序列一个是混沌问题, 另一个是实际问题, 因此能够体现其代表作用. 而且实际系统发生故障都是一个量变引起质变的过程, 所以本文选取指数函数作为故障信号.

3.1 Henon时间序列 (Henon time series)

选择Henon映射产生的300个数据作为时间序列进行仿真, 其中前250个数据作为训练集, 后50个数据作为测试集. 其产生方式如式(11)所示:

$$x_n=1-1.4\cdot x_{n-1}^2+0.3\cdot x_{n-2}. \tag{11}$$

取初值[0 0],系统的观测噪声为$N(0,0.05)$,故障信号可近似为$\exp(-(k-50)^2/50)$,表示第50步发生故障.

首先运用FPE准则估算嵌入维数, 由图2可知, 当$k=8$时, FPE取最小, 因此可以确定嵌入维数为8. 接着, 在训练数据中加入不同水平的噪声, 分别运用预先设定支持向量机超参数的方法和计算超参数的方法对Henon 时间序列进行预测, 具体结果如表1所示, 其中SVs表示支持向量的个数, Prio表示预先设定超参数($C=1000,\varepsilon=0.01,\sigma=0.5$), 而Selec表示通过考察时间序列数据计算支持向量机的超参数, 表1明显可以看出计算支持向量机超参数方法的优越性. 然后, 根据式(4)对Henon时间序列重构输入输出空间, 并基于支持向量机回归的预测算法建立预测模型. 接着, 对用于训练的Henon时间序列根据K-均值聚类算法进行聚类学习, 得到的正常时间序列原型. 然后, 运用预测模型, 根据式(8)的多步预测算法, 对测试序列进行时间序列的预测, 得到预测序列. 最后, 可以将预测序列与观测序列对比(图3(a))或计算预测序列与正常时间序列原型的距离(图3(b)), 实现故障预报.

计算预测序列与正常时间序列原型的距离进行故障预报, 相比于将预测序列与观测序列对比实现故障预报而言, 有两大优点. 其一, 更能够满足实时

性; 表面上看, 对比预测序列与观测序列的方法要比计算预测序列与正常时间序列原型的欧氏距离实现故障预报的方法更为迅速, 图3(a)中可以在第42步时预报故障，而图3(b)则要到第47步, 实则不然. 因为即使采用了多步预测技术, 前者一定要到第42步才能实现故障预报, 毕竟这一方法需要第42步的观测值. 后一种方法, 虽说在第47步实现故障预报, 实际上却是在第39步(嵌入维数是8). 其二, 预报结果更为准确. 若是系统在正常工作时, 出现离群点, 前一种方法会出现误报现象, 而本文提出的方法可以克服这种不利情况.

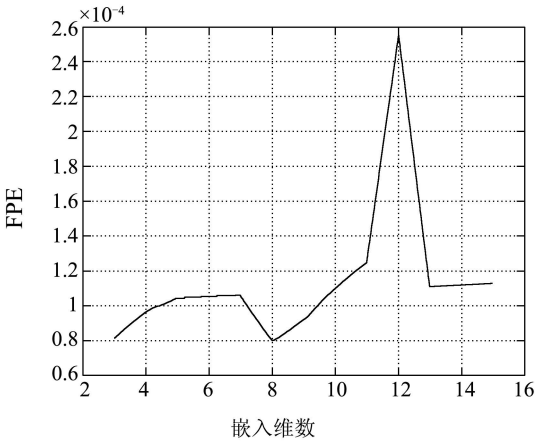


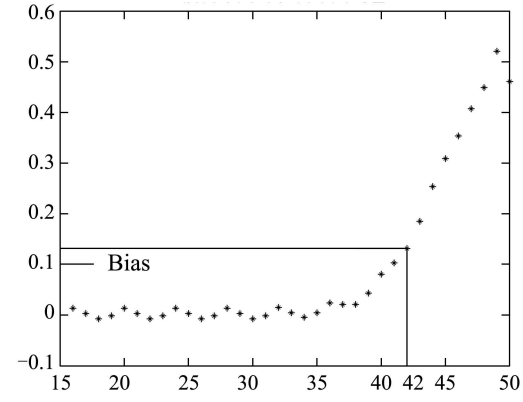
图 2　选择最佳嵌入维数

Fig. 2　Selecting the optimal embedding dimension

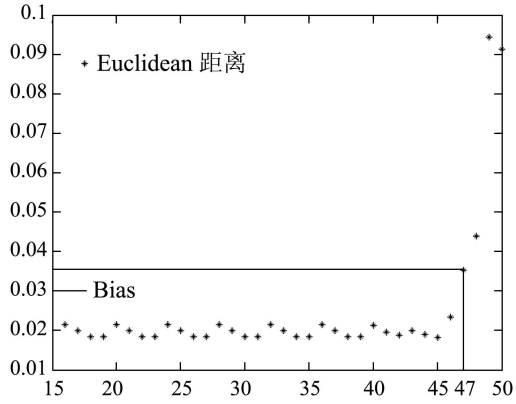
表 1　预先设定和计算超参数的预测效果对比

Table 1　Comparison between Prio and Selec

噪声类型	NMSE		SVs	
	Prio	Selec	Prio	Selec
N(0,0.005)	4.7×10^{-6}	4.7×10^{-6}	228(93.1%)	107(44.2%)
N(0,0.01)	6.2×10^{-5}	7.3×10^{-6}	233(96.3%)	135(55.8%)
N(0,0.05)	2.4×10^{-3}	1.1×10^{-3}	240(99.2%)	199(82.2%)



(a) 预测序列与观测序列之差



(b) 预测序列与正常原型的距离

图 3　基于Henon时间序列的故障预报

Fig. 3　Fault prediction based on Henon time series

3.2　釜式反应器 (CSTR)

一个连续搅拌釜式反应器[10]由式(12)(13)所示:

$$\frac{\mathrm{d}C_A}{\mathrm{d}t}=\frac{q}{V}(C_{Af}-C_A)-k_0\exp(-\frac{E}{RT})C_A, \quad (12)$$

$$\frac{\mathrm{d}T}{\mathrm{d}t}=\frac{q}{V}(T_f-T)+\frac{-\Delta H}{\rho C_p}k_0\exp(-\frac{E}{RT})C_A+\frac{UA}{V\rho C_p}(T_c-T). \quad (13)$$

式中: C_A是反应浓度, T是反应温度, T_c是冷却剂的温度, q是反应物进料流速, C_{Af}是进料流速, T_f是进料温度, V是反应釜的体积, ρ是密度, k_0是预指数因子, EH是活化能, $-\Delta H$是反应热, C_p是热容量. 在标称状态下, 此反应釜的参数参见文献[10].

采用欧拉离散法将系统离散化, 采样间隔选为: $\mathrm{d}t=0.2$ min, 初始状态为: $C_A(0)=0.2$ mol/L, $T(0)=400$ K. 反应器的控制律采用最简单的基于状态反馈的数值PID 控制律, 进行浓度设定点的跟踪控制. 对系统设置一个故障, 从第150步开始, 进料流速沿指数曲线下降:

$$q(k)=q(150)+1-\exp(k-150/65).$$

以反应温度为监控对象, 根据基于时间序列的故障预报算法, 确定最佳嵌入维数为8, 相应的支持向量机回归预测的步长为8, 聚类数目选为4. 图4为系统在正常和故障状态情况下的仿真曲线, 图5表示运用本文的方法实现故障预报, 由于各特征之间存在着相关性, 此时基于Mahalanobis距离度量的效果要明显优于Euclidean 距离度量, 并且在162步(170–8)预报系统即将发生故障.

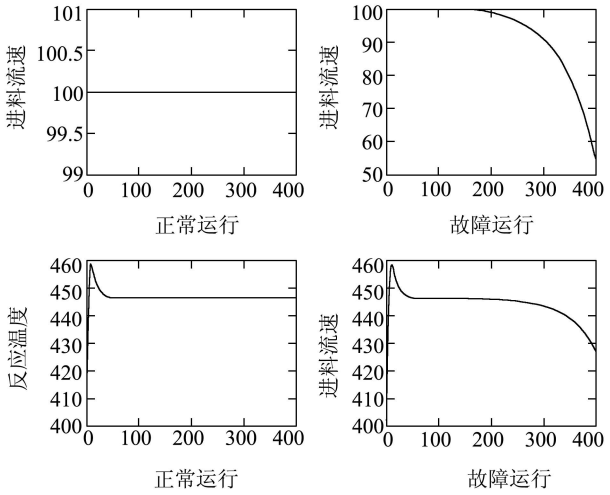


图 4 系统状态

Fig. 4 States of the CSTR

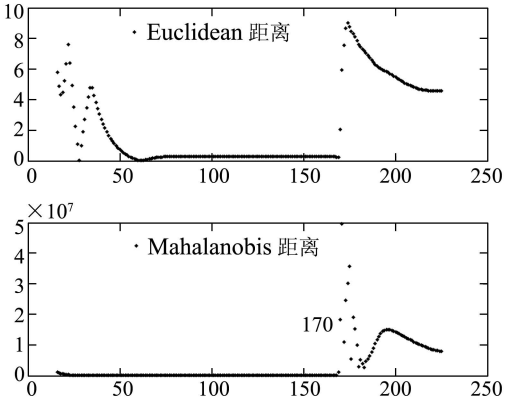


图 5 时间序列相似性度量

Fig. 5 Time series similarity measure

4 结论 (Conclusions)

本文提出的故障预报算法主要有3部分组成: 聚类、时间序列预测和计算时间序列相似度. 聚类的主要目的是得出正常时间序列的原型, 以便与预测序列对比实现故障预报. 至于利用支持向量机实现时间序列预测, 本文采取的计算支持向量机超参数的办法可以兼顾预测的准确性和泛化性能.

总之, 本文的方法较之以往单独比较预测值和观测值的做法更贴近故障预报的本质, 而且更为迅速和准确.

参考文献(References):

[1] HO S L, XIE M. The use of ARIMA models for reliability forecasting and analysis[J]. *Computer & Industrial Engineering*, 1998, 35(1-2): 213 – 216.

[2] TSE P W, ATHERTON D P. Prediction of machine deterioration using vibration based fault trends and recurrent neural networks[J]. *J of Vibration & Acoustics*, 1999, 121(7): 355 – 362.

[3] JAIN A K, MURTY M N. Data clustering: a review[J]. *ACM Computing Surveys*, 1999, 31(3): 264 – 323.

[4] RICHARD O D, PETER E H, DAVID G S. *Pattern Classification*[M]. Second Edition. New York: John Wiley & Sons, 2002.

[5] VAPNIK V N. *The Nature of Statistical Learning Theory*[M]. New York: Springer, 1995.

[6] THISSENA U, BRAKELA R, et al. Using support vector machines for time series prediction[J]. *Chemometrics & Intelligent Laboratory Systems*, 2003, 6(9): 35 – 49

[7] VAPNIK V N. *Statistical Learning Theory*[M]. New York: Wiley, 1998.

[8] JAMES T K, IVOR W T. Linear dependency between ε and the input noise in ε-support vector regression[J]. *IEEE Trans on Neural Networks*, 2003, 14(5): 544 – 553.

[9] MATTERA D, HAYKIN S. Support vector machines for dynamic reconstruction of a chaotic system[C]*// Advances in Kernel Methods-Support Vector Learning*. Cambridge: MIT Press, 1999: 243 – 254.

[10] 周东华, 叶银忠. 现代故障诊断与容错控制[M]. 北京: 清华大学出版社, 2000.
(ZHOU Donghua, YE Yinzhong. *Modern Fault Diagnosis and Fault Tolerant Control*[M]. Beijing: Tsinghua University Press, 2000.)

作者简介:

张军峰 (1979—), 男, 博士研究生, 主要研究方向为神经网络、故障检测和故障预报, E-mail: wufeng7919@163.com;

胡寿松 (1937—), 男, 教授, 博士生导师, 主要研究方向为故障诊断、自修复控制和复杂系统控制.