文章编号: 1000-8055(2006)04-0767-6

支持向量机时间序列预测模型的参数影响分析与自适应优化

杨虞微, 左洪福, 陈 果
(南京航空航天大学 民航学院,南京 210016)

摘 要: 建立在统计学习理论和结构风险最小原则上的支持向量机在理论上保证了模型的最大泛化能力,因此与建立在经验风险最小原则上的神经网络模型相比,理论上更为完善。本文运用支持向量机建立时间序列预测模型,研究影响模型预测精度的相关参数,在分析参数对时间序列预测精度的影响基础上,提出用遗传算法建立支持向量机预测模型的参数自适应优化算法。最后,用太阳黑子数据和航空发动机油样光谱数据进行了预测分析。算例表明了本文算法的正确性。

关 键 词: 航空、航天推进系统; 支持向量机; 时间序列分析; 预测; 遗传算法; 优化
中图分类号: O329; F201 **文献标识码**: A

Influence analysis and self-adaptive optimization of support vector machine time series forecasting model parameters

YANG Yu-wei, ZUO Hong-fu, CHEN Guo
(College of Civil Aviation,
Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract: Support Vector Machine (SVM) is based on Statistical Learning Theory (SLT) and Structural Risk Minimization Principle (SRM), and theoretically assures best generalization, therefore, it is theoretically better than Artificial Neural Network (ANN) which is based on Empirical Risk Minimization Principle (ERM). In this paper, SVM was used to establish time series forecasting model, and on the basis of analyzing the influence of model parameters, a self-adaptive optimizing algorithm based on genetic algorithm was put forward. Finally, the sunspot data and the spectrometric oil data of some aero-engines were used for preliminary analysis, and the results show the correctness and validity of the new method.

Key words: Aerospace propulsion system; Support Vector Machine (SVM); Time series analysis; Forecasting; Genetic algorithm; Optimization

时间序列分析是系统辨识和建模型的有效工具,在系统和系统的输入均未知的情况下,通过对系统输出的时间序列进行分析和建模,并在此基础上实现时间序列的外推预测。通常的时间序列分析方法是基于线性模型的 ARMA(n, m)法[1],由于通常模型均具有不同程度的非线性特征,特别对于非线性更为突出的复杂系统,ARMA 模型将产生较大的误差。由于神经网络能拟合任意的

收稿日期: 2005-07-30; **修订日期**: 2005-12-22

作者简介:杨虞微(1972-),男,南京航空航天大学民航学院载运工具运用工程博士研究生,主要从事航空器检测、诊断与维修领域的研究工作.

© 1994-2006 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

非线性函数并且具有一定的泛化能力,因此目前被广泛运用于时间序列预测[2,3]。但是,建立在经验风险最小原则基础上的神经网络模型,由于其泛化能力不能从理论上得到保证,从而导致了神经网络预测模型实际应用较为困难[4]。与之相比,由 V. Vapnik 创立的支持向量机[4]建立在统计学习理论和结构风险最小的基础上,在理论上充分保证了模型的泛化能力,具有较大的优势,目前已经广泛运用于时间序列分析和预测[5,6]。

但是,用于时间序列预测的支持向量机模型,本身也有许多参数要进行选择,比如嵌入维数、损失函数参数 ε,惩罚因子 C 以及核函数的相关参数等。这些参数在一定程度上对预测精度具有很大影响,且目前尚无统一选择标准。本文建立支持向量机的时间序列预测模型,在进行参数影响研究基础上,构造基于遗传算法的模型参数自适应优化算法。

1 时间序列预测原理

混沌理论[7]是研究非线性系统动力行为的新方法,它的基本观点是:简单的非线性系统可以产生简单的确定行为,也可以产生不稳定但有界的貌似随机的不确定现象,但这种随机不等同与统计学中的随机,本质上是复杂确定性系统产生的行为。由于混沌对初始条件极端敏感,这意味着混沌动力学特性能够放大微小的差异,导致宏观尺度上完全不可预测的程度,因此时间序列的长期预测不可行,只可能进行短期预测。

为了对非线性系统产生的时间序列进行预测,需要研究非线性系统的运动规律,把握其运动状态,这就要求从系统产生的时间序列中抽取动力系统,重构相空间,最常用的方法是时延法。

设所研究的时间序列为 $\{x(t)\}$, $t = 1, 2, \cdots, N$,则当前状态的信息可以表示成 m 维的延迟矢量:$x(t+\tau) = f\{x(t), x(t-\tau), \cdots, x(t-(m-1)\tau)\}$,式中,$m$ 为嵌入维数,τ 为时间延迟,通常取为采样间隔。Takens[8]已经证明:假设动力系统的维数为 d,如果 $m \geq 2d+1$,则这种映射产生的伪相空间和系统的状态空间微分同胚,及拓扑等价,他们的动力学特性定性意义上完全相同。

由此可见,对时间序列的预测,关键在于根据已知时间序列数据,对非线性系统相空间的重构,找出从 m 维空间映射到一维空间的映射函数。

由于多步预测可以由单步预测迭代而成,因此不失一般性,可以以单变量单步预测为例进行研究。设一个单变量时间序列 $\{x_1, x_2, \cdots\}$,对它进行预测的前提是认为其未来值与其前面的 m 个值之间有着某种函数关系,可描述为:

$$x_{n+k} = F(x_n, x_{n-1}, \cdots, x_{n-m+1}) \tag{1}$$

2 基于支持向量机的函数拟合

首先考虑线性回归问题,对于给定的训练样本 (x_i, y_i), $x_i \in R^d$, $y_i \in R$, $i = 1, \cdots, n$ 线性回归的目标就是求下列回归函数:

$$f(x) = \langle w, x \rangle + b \tag{2}$$

式中:$w \in R^d$; $b \in R$; $\langle w, x_i \rangle$ 为 w 与 x_i 的内积,并且满足结构风险最小化原理。对优化目标函数求极值:

$$Q(w) = \frac{1}{2}\langle w, w \rangle + CR_{emp}(f) \tag{3}$$

式中:C 为惩罚因子,实现在经验风险和置信范围之间的折中;$R_{emp}(f)$ 为损失函数,常用的损失函数有二次函数、Huber 函数、Laplace 函数和 ε 不敏感函数,其中 ε 不敏感函数可以确保对偶变量的稀疏性,同时确保全局最小解的存在和可靠泛化界的优化。因为这些较好的性质而得到广泛的应用,其定义为:

$$L_\varepsilon(d, y) = \begin{cases} |f(x) - y| - \varepsilon, & |f(x) - y| > \varepsilon \\ 0 & \text{其他} \end{cases} \tag{4}$$

当引入 ε 不敏感函数时,式(3)可写为:

$$Q(w) = \frac{1}{2}\langle w, w \rangle + C\frac{1}{n}\sum_{i}^{n} |y_i - f(x_i)|_\varepsilon \tag{5}$$

显然,当 $|y_i - \langle w, x_i \rangle - b| \leq \varepsilon (i = 1, 2, \cdots, n)$,即所有样本均落在由 $f(x) + \varepsilon$ 和 $f(x) - \varepsilon$ 组成的带状区域内时,优化问题就变为:

$$\min \frac{1}{2}\langle w, w \rangle \tag{6}$$

$$\text{s.t} \quad y_i + \langle w, x_i \rangle - b \leq \varepsilon,$$
$$\langle w, x_i \rangle - y_i + b \leq \varepsilon$$

考虑到上述条件不能充分满足,引入松弛因子 $\zeta \geq 0$ 和,$\zeta^* \geq 0$ 则式(6)的优化问题变为:

$$\min \frac{1}{2}\langle w, w \rangle + C\sum_{i}^{n}(\zeta_i + \zeta_i^*) \tag{7}$$

$$\text{s.t} \quad y_i - \langle w, x_i \rangle + b \leq \varepsilon + \zeta_i,$$
$$\langle w, x_i \rangle - y_i + b \leq \varepsilon + \zeta_i^*$$

上述问题可以通过求解最大化二次型的参数 α_i, α_i^* 而得到解决:

$$Q(\alpha, \alpha^*) = \sum_{i}^{n} y_i(\alpha - \alpha_i^*) - \varepsilon\sum_{i}^{n}(\alpha_i + \alpha_i^*) - \frac{1}{2}\sum_{i=1, j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j \rangle \tag{8}$$

© 1994-2006 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$\text{s.t} \quad \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0, 0 \le \alpha_i \le C,$$

$$i = 1,2,\cdots,n \quad 0 \le \alpha_i^* \le C, i = 1,2,\cdots,n$$

式中:求解出上述各参数 α_i,α_i^* 后,就可以利用:

$$b = -1/2 \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)(< x_i, x_t > + < x_i, x_s >) \tag{9}$$

求得 b,其中,x_s,x_t 为任选的两个非支持向量。这样就可得到拟合函数:

$$f(\alpha,\alpha^*, x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) < x, x_i > + b \tag{10}$$

用核函数 $K(x_i, x_j)$ 来替代内积运算,实现由低维空间到高维空间的映射,从而使低维空间的非线性问题转化为高维空间的线性问题。引入核函数后,优化目标函数式(8)变为如下形式:

$$Q(\alpha,\alpha^*) = \sum_{i=1}^{n} y_i(\alpha_i - \alpha_i^*) - \varepsilon\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) - \frac{1}{2}\sum_{i=1,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(x_i, x_j) \tag{11}$$

而相应的拟合函数式(10)也变为:

$$f(\alpha,\alpha^*, x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) K(x, x_i) + b \tag{12}$$

进行时间序列分析通常要建立自回归模型,它是一个动态模型,当前时刻的值与以前 $n-1$ 个时刻的值均有关系,即需要建立输入向量 $x_t = \{x_{t-1}, x_{t-2}, \cdots, x_{t-p}\}$ 与输出 $\{x_t\}$ 之间建立一一映射关系:$f: R^p \to R$,其中 p 为嵌入维数。根据以上思想形成训练样本集:

$$X = \begin{vmatrix} x_1 & x_2 & \cdots & x_{n-p} \\ x_2 & x_3 & \cdots & x_{n-p+1} \\ \cdots & \cdots & O & \cdots \\ x_p & x_{p+1} & \cdots & x_{n-1} \end{vmatrix}, Y = [x_{p+1} \quad x_{p+2} \quad \cdots \quad x_n] \tag{13}$$

利用公式(13)建立时间序列预测模型为:

$$y_t = \sum_{i=1}^{n-p}(\alpha_i - \alpha_i^*) k(x_i, x_t) + b, t = p+1, \cdots, n \tag{14}$$

3　支持向量机预测模型参数对时间序列预测精度的影响分析

3.1　时间序列预测精度评价函数

在实际应用中,对于实际测得的时间序列 $\{x_1, x_2, \cdots\}$,可以利用其一部分数据来建模,而用另一部分数据来对所建模型进行验证,如果预测值是实测值相差越少,显然模型越理想,理想情况是预测值与实测值相等,则达到完美预测。通常衡量预测值与实测值差别的变量采用平均相对变动值(Average Relative Variance:ARV)[9],其定义为:

$$ARV = \sum_{i=1}^{N}[x(i) - \hat{x}(i)]^2 / \sum_{i=1}^{N}[x(i) - \bar{x}(i)]^2 \tag{15}$$

其中:N - 比较数据个数;$x(i)$ - 实测数据值;$\bar{x}$ - 实测数据平均值;$\hat{x}(i)$ - 预测值。显然,平均相对变动值 ARV 越小,也表明预测效果越好,ARV = 0 表示达到了理想预测效果,当 ARV = 1 时,表明模型仅达到平均值的预测效果。

显然,当 N 为整个时间序列的长度时,ARV 将综合反映了 SVM 预测模型对训练点和预测点的拟合程度。

3.2　评价神经网络预测精度的时间序列

为了分析神经网络结构对时间序列预测精度的影响,需要选取标准的时间序列来进行实验。国际上所采用的标准时间序列有很多,其中太阳黑子数据是最具有代表性的数据之一。本文选取 1700 至 1987 年的太阳黑子数据,其中用前 1/5 数据进行 SVM 建模,用后 4/5 数据来对模型进行检验。

3.3　支持向量机预测模型参数的分析

支持向量机是建立在统计学习坚实的理论基础之上的,具有理论的完备性,但是在应用上,仍然存在一些问题,典型的问题就是模型参数的选择,对预测精度有重要影响的参数是:嵌入维数 p、损失函数参数 ε,惩罚因子 C,核函数及其参数的选取。

(1)嵌入维数 p:关系到能否重构非线性系统的相空间。对时间序列预测精度有重要影响。

(2)损失函数的参数 ε 控制回归逼近误差管道的大小,从而控制支持向量的个数和泛化能力,其值越大,精度越低,则支持向量越少。ε 的取值范围一般为 0.000 1～0.1。

(3)惩罚因子 C 用于控制模型复杂度和逼近误差的折中,C 越大则对数据的拟合程度越高,但泛化能力越差,一般为 1～1 000 000。

(4)对不同的类型的核函数,所产生的支持向量的个数变化不大,但是核函数的相关参数,如对于多项式核函数,其多项式次数(一般为 2～9);对于径向基核函数,其 σ 值(一般为 0.1～3.8),对模型的预测精度有重要影响。本论文选定多项

© 1994-2006 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

式核函数,对其次数 Np 进行优化研究。

因此,尽管支持向量机在理论上很完善,但是在具体使用中,仍然存在模型参数的优化问题,目前,也无统一的模型选取标准和理论。表 1～表 4 分别为同一时间序列在不同的模型参数下所得到的预测精度比较:

表 1 惩罚因子 C 对预测精度的影响

Table 1 The influence of penalty factor C on forecasting precision

惩罚因子 C	损失函数参数 ε	多项式核函数次数 Np	嵌入维数 p	平均相对变动值 ARV
1	0.01	2	9	0.2320
10	0.01	2	9	0.7341
100	0.01	2	9	163.32
1000	0.01	2	9	17820
10000	0.01	2	9	79785

表 2 损失函数参数 ε 对预测精度的影响

Table 2 The influence of loss function parameter ε on forecasting precision

惩罚因子 C	损失函数参数 ε	多项式核函数次数 Np	嵌入维数 p	平均相对变动值 ARV
1	0.1	2	9	0.2040
1	0.01	2	9	0.2320
1	0.001	2	9	0.2450
1	0.0001	2	9	0.2451
1	0.00001	2	9	0.2450

表 3 多项式核函数次数 Np 对预测精度的影响

Table 3 The influence of polynomial kernel function order Np on forecasting precision

惩罚因子 C	损失函数参数 ε	多项式核函数次数 Np	嵌入维数 p	平均相对变动值 ARV
1	0.01	1	9	0.2534
1	0.01	2	9	0.2320
1	0.01	3	9	0.4660
1	0.01	4	9	0.9147
1	0.01	5	9	1.5147

表 4 嵌入维数 p 对预测精度的影响

Table 4 The influence of embedded dimension p on forecasting precision

惩罚因子 C	损失函数参数 ε	多项式核函数次数 Np	嵌入维数 p	平均相对变动值 ARV
1	0.01	2	1	0.5285
1	0.01	2	3	0.2165
1	0.01	2	5	0.3902
1	0.01	2	7	0.4134
1	0.01	2	9	0.2320

通过比较表 1 至表 4,可以看出:除损失函数参数 e 的影响相对较小外,其他参数对预测结果的影响均很大。图 1 为参数 $C=1$, $\varepsilon=0.1$, $Np=2$, $p=9$ 时实测值与预测值的比较,其中预测误差 ARV = 0.204 0,为所有计算中预测效果最好的。而图 2 为参数 $C=1$, $\varepsilon=0.01$, $Np=5$, $p=9$ 时实测值与预测值的比较,其中预测误差 ARV = 1.514 7,为所有计算中预测效果较差的。从图 1 和图 2 的比较,不难看出支持向量机预测模型参数对其预测精度的影响程度。

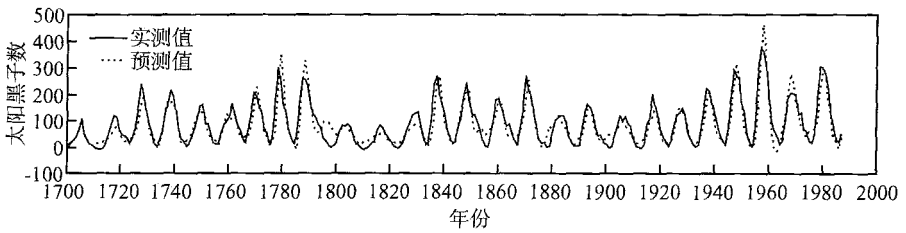

图 1 参数 $C=1$, $\varepsilon=0.1$, $Np=2$, $p=9$ 时实测值与预测值的比较

Fig. 1 The comparison of measured and forecasted valued under $C=1$, $\varepsilon=0.1$, $Np=2$, $p=9$

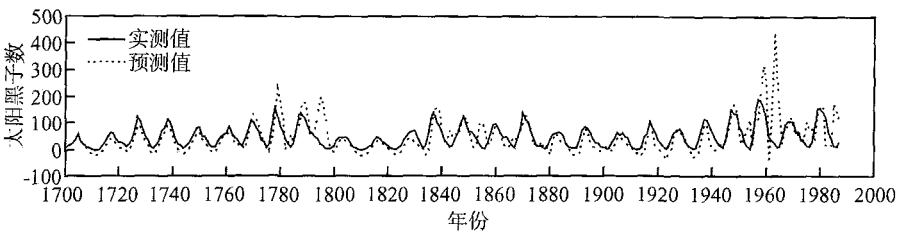

图 2 参数 $C=1$, $\varepsilon=0.01$, $Np=5$, $p=9$ 时实测值与预测值的比较

Fig. 2 The comparison of measured and forecasted valued under $C=1$, $\varepsilon=0.01$, $Np=5$, $p=9$

© 1994-2006 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

4　支持向量机预测模型参数优化的自适应算法

通过上述分析发现,支持向量机模型预测精度与惩罚因子 C、损失函数参数 ε、多项式核函数次数 Np 及嵌入维数 p 均存在一定的关系,为了获取最佳预测性能的 SVM 模型,需要得到最佳的 C,ε,Np 和 p 值。显然这是一个优化问题,如果采取穷举的方式搜索最优值,计算量将十分巨大以至于无法实现。由于遗传算法[10]具有隐含的并行性和强大全局搜索能力,可以在很短的时间内搜索到全局最优点。

因此。本文利用遗传算法来进行 SVM 预测模型的参数优化。首先,对 SVM 预测模型惩罚因子 C、损失函数参数 ε、多项式核函数次数 Np 及嵌入维数 p 进行二进制编码,并随机产生初始化种群。其次,对种群中的各染色体解码,获取 C,ε,Np 及 p 值,运用一部分数据建立 SVM 预测模型,计算所有数据的预测值与实测值的 ARV 值,从而得到各基因串的适应度。然后判断遗传算法的停止准则是否满足,如果满足则停止计算,输出最优参数,否则,则执行选择、交叉和变异等操作以产生新一代种群,并开始新一代的遗传。在本文中,用达到给定的遗传代数作为计算的停止条件。

本文遗传算法中:交叉率和变异率分别为 0.50 和 0.05。基因串(染色体)中 C,ε,Np 和 p 值均采用二进制编码,其染色体结构如图 3 所示。染色体中的排列顺序为 p,Np,C 及 e,设它们的位数分别为 n_1,n_2,n_3 及 n_4,则染色体的长度为 $n_1+n_2+n_3+n_4$,搜索空间为 $2^{n_1+n_2+n_3+n_4}$。

为了避免嵌入维数和多项式次数为 0,规定解码后,加上 1 得到嵌入维数 p 和多项式次数 Np,由于 C 和 ε 均为实数,对于 C,通过计算 $C=10^C$ 得到惩罚因子 C,对于 ε,通过计算 $\varepsilon=0.1^{\varepsilon}$ 得到损失函数参数 ε。

由于实测值与预测值的平均相对变动值 ARV 充分反映了模型的预测精度,因此,将遗传算法的适应度函数取为 $f=1/ARV$。

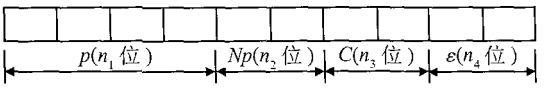

图 3　染色体结构

Fig. 3　Chromosome structure

下面列举两个算例以表明本文算法的有效性:

(1) 算例一:太阳黑子数据

用前 20 %的数据建模,后 80 %的数据验证。遗传算法的计算参数为:交叉率和变异率分别为 0.50 和 0.05。种群数为 10,进化代数为 10。参数 p,Np,C 及 ε 的二进制编码长度分别为 $n_1=4$、$n_2=2$、$n_3=2$ 及 $n_4=2$。通过 10 代遗传后得到了最优的参数:$p=9$, $Np=2$, $C=1$, $\varepsilon=0.01$。最优的适应度值为 9.689 6,平均相对变动值 ARV 为 0.232 0。图 4 为该模型对太阳黑子数据的预测值与实测值的比较。从图中可以看出其拟合程度达到很好的拟合效果。

(2) 算例二:航空发动机油样光谱数据

为了进一步验证本文算法的有效性,选取某航空发动机的光谱分析数据作为算例,光谱分析仪器为美国 Bird 公司的原子发射光谱分析仪。用该数据的前 50 %建模,后 50 %验证。遗传算法的计算参数为:交叉率和变异率分别为 0.50 和 0.05。种群数为 30,进化代数为 10。参数 p,Np,C 及 ε 的二进制编码长度分别为 $n_1=2$, $n_2=2$, $n_3=2$ 及 $n_4=2$。通过 10 代遗传后得到了最优的参数:$p=4$, $Np=1$, $C=1$, $\varepsilon=0.001$。最优的适应度值为 3.250 5,平均相对变动值 ARV 为 0.307 6。图 5 为该模型的预测值与实测值比较。从图中可以看出其拟合程度也达到很好的拟合效果。

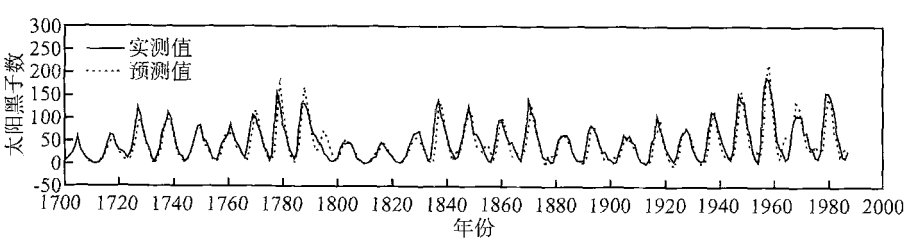

图 4　自适应优化算法获取的最优 SVM 模型对太阳黑子数据的预测结果

Fig. 4　The forecasting result of sunspot data of optimum SVM model obtained by self - adaptive Algorithm

© 1994-2006 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

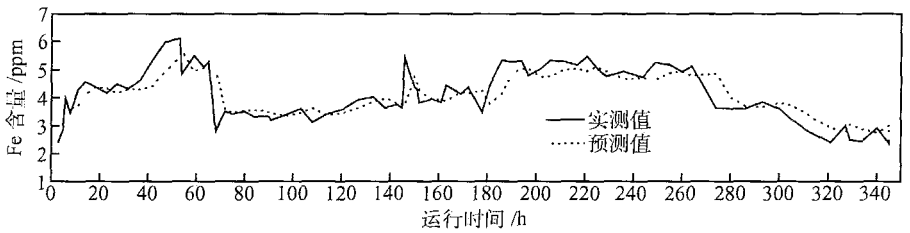

图5 自适应优化算法获取的最优 SVM 模型对光谱油样数据的预测结果

Fig. 5 The forecasting result of Spectral oil data of optimum SVM model obtained by self - adaptive Algorithm

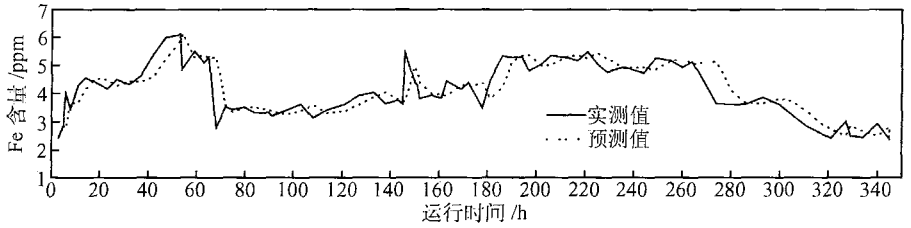

图6 ARMA(2,0)模型对光谱油样数据的预测结果

Fig. 6 The forecasting result of Spectrometric oil data of ARMA(2,0) model

为了与 ARMA 模型进行比较,图6为建立 ARMA(2,0)模型,针对相同的光谱油样分析数据,同样用前50%建模,后50%进行外推预测得到的实测值与预测值的比较结果。其实测值与预测值的平均相对变动值 ARV 为0.367 8,显然预测效果较本文算法差。该比较结果在一定程度上进一步验证了本文基于遗传算法的支持向量机预测模型参数优化算法的有效性。

5 结 论

(1)分析了运用支持向量机进行非线性预测的优越性以及存在的问题;

(2)提出了影响 SVM 预测能力的4个重要参数 - 嵌入维数 p、多项式核函数次数 Np、惩罚因子 C 及损失函数参数 ε。并对4个参数进行了预测精度影响分析;

(3)本文用遗传算法构造了同时优化影响 SVM 预测精度的参数(嵌入维数 p、多项式核函数次数 Np、惩罚因子 C 及损失函数参数 ε)的算法,利用遗传算法自动获取最优的 SVM 预测模型。最后用算例表明了本文算法的正确有效性。

参考文献:

[1] Lapedes A, Farber. Nonlinear signal processing using neural network: Prediction and system modeling [R]. *Technical Report LA-UR-87-2662, Los Alamos National Laboratory. Los Alamos. NM*, 1987.

[2] Weigend A, Rumelhart De, Huberman B A. Predicting the future: a connectionist approach [J]. *International Journal of Neural System*, 1990, (1):195-220.

[3] Vapnik V. The nature of statistical iearning [M]. *New York: Springer*, 1995.

[4] Francis E H Tay, Lijuan Cao. Application of support vector machines in financial time series forecasting [J]. *Omega*, 2001, 29: 309-317.

[5] 尉询楷,李应红,王硕,等. 基于支持向量机的航空发动机滑油监控分析[J]. 航空动力学报, 2004, 19(3):392-397.
Wei Xunkai, Li Yinghong, Wang Shuo, et al. Aero-engine lubrication monitoring analysis via support vector machines [J]. *Journal of Aerospace Power*, 2004, 19(3):392-397.

[6] Ford J. Chaos at random[J]. *Nature*, 1983, 305(20): 17-24.

[7] Takens F. Detecting strange attractors in turbulence[A]. In: Rand D A, Young L S. Dynamical Systems and Turbulence[C]. *Berlin: Springer-Verlag*, 1981.

[8] Cholewo T, Zurada J M. Sequential network construction for time series prediction [A]. Proceedings of the IEEE International Joint Conference on Neural Networks [C]. 1997: 2034-2039.

[9] Goldberg D. Genetic algorithms in search, optimization and machine learning [M]. *Addison-Wesley, Reading, MA*, 1989.

© 1994-2006 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net