基于遗传算法的支持向量机时间序列预测模型优化*

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,用算例表明了本文算法的正确性和有效性。

关键词 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 **国家标准学科分类代码** 110.6760

Optimizing of support vector machine time series forecasting model parameters based on genetic algorithms

Chen Guo
(Civil Aviation College, 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 model generalization. Therefore, it is more perfect in theory than Artificial Neural Network (ANN) that is based on Empirical Risk Minimization Principle (ERM). In this paper, SVM is used to establish time series forecasting model, study the parameters that influence forecasting accuracy. On the basis of analyzing model parameters' influence, a self-adaptive optimizing algorithm for establishing the model parameters based on genetic algorithm is put forward. In the end, examples showing the correctness and validity of the proposed algorithm are given.

Key words support vector machine (SVM) time series analysis forecasting genetic algorithm optimizing

1 引 言

时间序列分析是系统辨识和建模的有效工具,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和平稳时间序列的 ARMA(n, m)法[1],由于通常模型均具有不同程度的非线性特征,特别是复杂系统,其非线性特征更为突出,因此 ARMA 模型往往失效。由于神经网络能拟合任意的非线性函数并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,神经网络学习的不确定性和泛化能力不能保证等问题,导致了神经网络的结构设计困难等问题[4]。与建立在经验风险最小原则基础上的神经网络模型相比,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,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

* 本文于 2005 年 7 月收到。

自适应优化算法。

2　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,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,它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:

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

本文将研究用支持向量机来拟合式(1)中的函数 F。

3　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),$x\in R^d$,$y_i\in R$,$i=1,\cdots,n$,线性回归的目标就是求下列回归函数:

$$f(x)=(w\cdot x)+b \tag{2}$$

式中:$w\in R^d$;$b\in R$;$(w\cdot x)$ 为 w 与 x 的内积,且满足结构风险最小化原理。

对优化目标函数求极值:

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

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:

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

当引入 ε- 不敏感函数时,式(3)可写为:

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

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

$$\min\frac{1}{2}(w\cdot w)$$
$$\text{s.t}\quad y_i+(w\cdot x_i)-b\leq\varepsilon,\ (w\cdot x_i)-y_i+b\leq\varepsilon \tag{6}$$

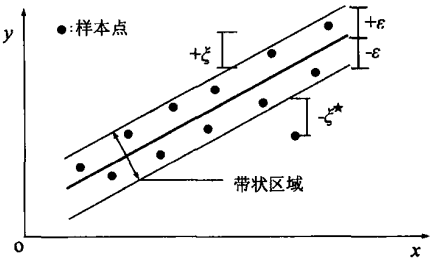

图 1　线性回归的不敏感区域

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:

$$\min\frac{1}{2}(w\cdot w)+C\sum_{i}^{n}(\zeta_i+\zeta_i^*)$$
$$\text{s.t}\quad y_i-(w\cdot x_i)+b\leq\varepsilon+\zeta_i,\ (w\cdot x_i)-y_i+b\leq\varepsilon+\zeta_i^* \tag{7}$$

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:

$$Q(\alpha,\alpha^*)=\sum_{i}^{n}y_i(\alpha_i-\alpha_i^*)-\varepsilon\sum_{i}^{n}(\alpha_i+\alpha_i^*)-\frac{1}{2}\sum_{i=1}^{n}\sum_{j}^{n}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)(x_i\cdot x_j)$$
$$\text{s.t}\quad\sum_{i}^{n}(\alpha_i-\alpha_i^*)=0,\ 0\leq\alpha_i\leq C,i=1,2,\cdots,n,0\leq\alpha_i^*\leq C,i=1,2,\cdots,n \tag{8}$$

求解出上述各参数 α_i、α_i^* 后,就可以利用:

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

求得 b,其中,x_s、x_t 为任选的 2 个非支持向量。

这样就可得到拟合函数:

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

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:

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

而相应的拟合函数式(10)也变为:

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

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:

$$\left.\begin{aligned}X&=\begin{bmatrix}x_1 & x_2 & \cdots & x_{n-p}\\ x_2 & x_3 & \cdots & x_{n-p+1}\\ \cdots & \cdots & & \cdots\\ x_p & x_{p+1} & \cdots & x_{n-1}\end{bmatrix}\\ Y&=\begin{Bmatrix}x_{p+1} & x_{p+2} & \cdots & x_n\end{Bmatrix}=\begin{Bmatrix}y_{p+1} & y_{p+2} & \cdots & y_n\end{Bmatrix}\end{aligned}\right\} \tag{13}$$

利用式(13)建立时间序列预测模型为:

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

4 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

4.1 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$\{x_1,x_2,\cdots\}$,可以利用其中一部分数据来建模,而用另一部分数据来对所建模型进行验证,如果预测值与实测值相差越少,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:

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

式中:N 为比较数据个数,$x(i)$ 为实测数据值,$\bar{x}$ 为实测数据平均值,$\hat{x}(i)$ 为预测值。显然,平均相对变动值 ARV 越小,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。

4.2 评价预测精度的时间序列

为了分析 SVM 模型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 1700～1987 年的太阳黑子数据,其中用前 1/5 数据进行 SVM 建模,用剩下的数据来对模型进行检验。

4.3 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。C 的取值范围一般为(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。本论文选定多项式核函数,对其次数 Np 进行优化研究。

表 1～4 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。通过比较表 1～4,可以看出,除损失函数参数 ε 的影响相对较小外,其他参数对预测结果的影响均很大。

表 1 惩罚因子 C 对预测精度的影响

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	0.232 0
10	0.01	2	9	0.734 1
100	0.01	2	9	163.32
1 000	0.01	2	9	17 820
10 000	0.01	2	9	79 785

表 2 损失函数参数 ε 对预测精度的影响

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	0.204 0
1	0.01	2	9	0.232 0
1	0.001	2	9	0.245 0
1	0.000 1	2	9	0.245 1
1	0.000 01	2	9	0.245 0

表 3 多项式核函数次数 Np 对预测精度的影响

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	0.253 4
1	0.01	2	9	0.232 0
1	0.01	3	9	0.466 0
1	0.01	4	9	0.914 7
1	0.01	5	9	1.514 7

表 4　嵌入维数 p 对预测精度的影响

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	0.528 5
1	0.01	2	3	0.216 5
1	0.01	2	5	0.390 2
1	0.01	2	7	0.413 4
1	0.01	2	9	0.232 0

5　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。

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。染色体中的排列顺序为 p、Np、C 及 ε,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 $2^{n1+n2+n3+n4}$。

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 $f = 1/ARV$。

6　算　　例

(1) 算例一:太阳黑子数据

用前 20 %的数据建模。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 2 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。

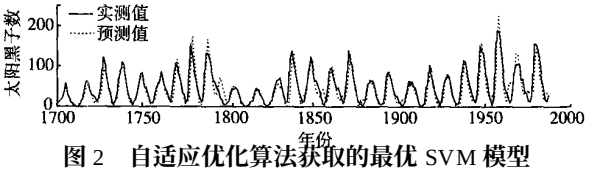

图 2　自适应优化算法获取的最优 SVM 模型对太阳黑子数据的预测结果

(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,用某航空发动机油样光谱数据进行实验,用前 50 %的数据建模。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 3 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。

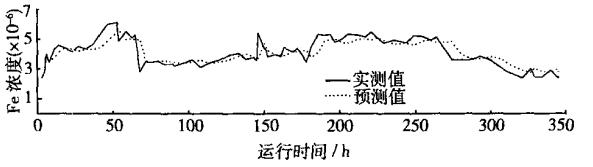

图 3　自适应优化算法获取的最优 SVM 模型对光谱油样数据的预测结果

7　结　　论

本文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;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。

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性和有效性。

参考文献

[1] 杨叔子, 吴雅. 时间序列分析的工程应用[M]. 武汉: 华中理工大学出版社, 1991.

[2] FARBER L A. 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.

[3] WEIGEND A B. Predicting the future: a connectionist approach [J]. International Journal of Neural System, 1990(1):193-209.

[4] VAPNIK V. The nature of statistical learning [M]. New York: Springer, 1995.

[5] TAY F E H, Cao L J. Application of support vector machines in financial time series forecasting [J]. Omega, 2001, 29: 309-317.

[6] 尉询凯,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, 2003, 18(6): 393-397.

[7] FORD J. Chaos at random[J]. Natrue, 1983, 305(20): 17-24.

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

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

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

作者简介

陈果 男 1972 年出生 副教授 主要研究方向为航空发动机状态监测与故障智能诊断 专家系统 数据融合 神经网络与遗传算法 图像处理及模式识别 机械动力学。

E-mail:cgzyx@263.net