

基于信息冗余检验的支持向量机时间序列预测 自由参数选取方法*

于艳华[†] 宋俊德

(北京邮电大学计算机学院, 北京 100876)

(2012年3月18日收到; 2012年6月13日收到修改稿)

支持向量机建模中的一个关键和难点问题自由参数的设置. 不同于以往应用残差的简单统计量选取最佳模型的方法, 本文提出通过检验模型在训练集上的拟合残差是否不含冗余信息作为选择自由参数的依据. 进一步提出应用全向相关函数 (omni-directional correlaton function, ODCF) 检验残差信息冗余并给出应用方法, 并从理论分析和数值仿真两方面给出该方法正确性的证明. 在两个典型的非线性时间序列 (年均太阳黑子数和 Mackey-Glass 数据) 上进行了实验, 实验结果优于相关文献记载及基于校验集方法的预测性能.

关键词: 支持向量机, 自由参数, 残差, 全向相关函数

PACS: 05.45.Tp, 05.45.-a, 02.50.Ey

1 引言

时间序列预测是数据挖掘的重要分支, 广泛应用于气候天气预测、财经金融市场预测、电力负荷预测、通信网络业务量预测等方面, 时间序列数据根据其特点可分为线性和非线性的、平稳和非平稳的、混沌和非混沌的等, 不同特性的序列需要不同的预测方法. 对于线性时间序列, Box-Jenkins 的 ARMA (auto-regressive moving average) 及其扩展的能处理含季节趋势项的 SARIMA (seasonal auto-regressive integrated moving average) 模型已经得到成熟应用^[1], 但自然界和人类活动产生的大量时间序列都是非线性的, 比如太阳黑子活动数据、经济活动中的股票序列、人民币美元汇率变动等序列已经证明是非线性甚至是混沌的, 对于广泛存在的非线性时间序列, 尤其是现在得到广泛研究和应用的混沌时间序列, 需要人工神经网络模型、支持

向量机等来预测^[2-7].

神经网络方法已经证明能以任意精度逼近任一函数, 在非线性系统预测中得到广泛应用. 但神经网络是基于传统的 Fisher 统计学中的经验风险最小化 (empirical risk minimization, ERM) 原则的, 追求的是训练集上的损失最小化, 因为训练数据的实际有限性, 不可避免地会出现过拟合的问题, 另外还存在神经网络结构需人为确定、可能收敛到局部极小等问题. 上世纪 90 年代中期, Vapnik 及其同事提出的支持向量机 (support vector machine, SVM) 解决了神经网络的以上问题^[8,9], 因此, 支持向量机在分类、回归以及非线性尤其是混沌系统时间序列预测领域得到越来越多的研究和应用^[2-7,10-14].

在支持向量机的研究和应用中, 一个问题始终没有完善解决: 支持向量机自由参数的选取问题^[10]. 支持向量机自由参数选取直接决定了所得支持向量从而决定了所得模型, 因此自由参数是否

* 国家自然科学基金 (批准号: 61072060), 国家科技支撑计划 (批准号: 2009BAH39B03), 国家高技术研究发展计划 (批准号: 2011AA100706), 高等学校博士学科点专项科研基金 (批准号: 20110005120007), 中央高校基本科研业务费专项资金 (批准号: 2012RC0205), 北京市教育委员会共建项目专项基金和教育部信息网络工程研究中心资助的课题.

[†] E-mail: yuyanhua@bupt.edu.cn

恰当直接决定了模型的好坏. 关于如何选取最优自由参数目前为止也有很多研究^[15-22], 但没有一个公认的最优选取方法; 应用较多的交叉验证方法非常耗时外也缺乏理论依据. 随着混沌理论研究的不断深入及其在信号处理、自动控制和通信领域中的广泛应用, 支持向量机在混沌时间序列预测中得到很大的关注和研究, 也取得很好应用效果^[2-7], 但在如何选取自由参数方面缺乏深入系统研究. 对于支持向量机应用于时间序列预测时更没有针对其不同于一般回归问题的特性的考虑: 一般的回归问题首先假定各样本点间相独立, 而时间序列各时间点观察值之间存在着相关性. 实际上时间序列观察值之间存在相关性正是时间序列建模和预测得以进行的基础; 时间序列建模的目的就是尽可能地提取这种相关性. 时间序列是一种动力学系统, 在动力学系统的系统辨识中, 一个重要步骤是检验所辨识模型的有效性. 基于以上分析, 论文提出一种新的 SVM 时间序列预测自由参数选取方法: 首先对训练残差进行信息冗余检验来确定模型的有效性, 而使得模型有效的自由参数是最优的. 另外, 考虑到时间序列的相关性可能是线性的也可能是非线性, 论文提出采用全向相关函数 (omni-directional correlation function, ODCF) 检验模型有效性, ODCF 可以全面检验线性非线性相关外, 还可以适用于各种容量的样本. 论文还给出了应用全向相关函数检验残差信息冗余的方法, 并从理论分析和数据仿真两方面给出证明.

2 支持向量机概述

古典统计学是建立在样本趋近无限多的渐进理论基础上的, 然而在实际中所能得到的样本总是有限的; 这就是为什么神经网络模型会出现过拟合的原因. 针对这个问题, Vapnik 等提出了统计学习理论 (statistical learning theory, SLT), 该理论除了考虑训练集上的损失, 还致力于使模型的复杂度尽可能小, 该机器学习原则称为结构风险最小化原则 (structural risk minimization, SRM). 而支持向量机就是建立在结构风险最小化原则基础上, 被视为替代神经网络的较好方法^[8,9]. SVM 作为一种机器学习方法, 可用于分类和回归 (函数逼近)^[10-14], 下面对支持向量回归机的原理作一简单介绍.

所谓回归问题, 是指通过用一个曲线逼近一个数据点集 (\mathbf{x}_i, y_i) ($\mathbf{x}_i \in \mathbb{R}^m, y_i \in \mathbb{R}, i = 1, \dots, l$) 来估计一个函数 $f(\mathbf{x})$, 使得下式成立:

$$y_i = f(\mathbf{x}_i) + e_i, \quad (1)$$

其中, e_i 为均值为 0 方差为 σ^2 的噪声. 支持向量回归机则是支持向量机在回归领域的应用. 目标函数 $f(\mathbf{x})$ 可能是线性也可能是非线性的, 不妨统一表述为

$$f(\mathbf{x}) = (\mathbf{w} \cdot \phi(\mathbf{x})) + b, \quad (2)$$

在 (2) 式中, 当 $\phi(\cdot)$ 为非线性函数时, $f(\mathbf{x})$ 在输入空间是非线性的; 如果把函数 $\phi(\cdot)$ 看作从输入空间到特征空间 F 的映射函数, 则函数 $f(\mathbf{x})$ 可以看作特征空间 F 中的线性函数, $F = \{\phi(\mathbf{x}) | \mathbf{x} \in \mathbf{X}\}$. 在该式中, $\mathbf{w} \in F$ 是系数向量, $b \in \mathbb{R}$ 是截矩 (偏置).

作为基于 SRM 原则的机器学习方法, SVM 通过求解以下的不等式约束优化问题构建回归模型:

$$\begin{aligned} & \min_{\mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}, \xi_i^* \in \mathbb{R}^{2l}} R(\mathbf{w}, \xi, \xi^*) \\ & = \frac{1}{2} \|\mathbf{w}\|^2 + C \cdot \frac{1}{l} \sum_{i=1}^l (\xi_i + \xi_i^*), \\ & \text{s.t. } ((\mathbf{w} \cdot \phi(\mathbf{x}_i)) + b) - y_i \leq \varepsilon + \xi_i, \\ & \quad i = 1, 2, \dots, l, \\ & \quad y_i - ((\mathbf{w} \cdot \phi(\mathbf{x}_i)) + b) \leq \varepsilon + \xi_i^*, \\ & \quad i = 1, 2, \dots, l, \\ & \quad \xi_i \geq 0 \quad i = 1, 2, \dots, l, \\ & \quad \xi_i^* \geq 0 \quad i = 1, 2, \dots, l, \end{aligned} \quad (3)$$

可以看出, 该优化问题的目标函数分为两部分: $\frac{1}{l} \sum_{i=1}^l (\xi_i + \xi_i^*)$ 是采用 ε - 不敏感损失函数时训练集上经验损失; $\frac{1}{2} \|\mathbf{w}\|^2$ 是指示函数复杂度的 VC(Vapnik-Chervonenkis) 维的界. ε - 不敏感损失函数为 $\text{loss}(x) = \max(|f(x) - y| - \varepsilon, 0)$, $\varepsilon > 0$. 式 (3) 中有两个参数需要在解优化问题前作为优化问题常数人为设定: C 和 ε . 其中, ε 决定了误差敏感区域的宽度; C 称为惩罚 (正则) 参数, 用于调节训练集上损失值与决策函数的复杂度两部分的比重. 这种无法由学习得到, 而需在学习前人为设定的参数称为自由参数 (free parameter, 又称元参数 meta-parameter 或超参数 hyper-parameter)^[10]. 除

了这两个自由参数外,支持向量机还需要设定核函数相关自由参数,而核函数的引入则与求解(3)式优化问题的过程有关.

为了求解(3)式的优化问题,应用 Lagrange 对偶原理,则原优化问题可以转化为以下更容易求解的优化问题:

$$\begin{aligned} \min & \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l (\alpha_i^* - \alpha_i) (\alpha_j^* - \alpha_j) \\ & \times (\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)) \\ & + \varepsilon \sum_{i=1}^l (\alpha_i^* + \alpha_i) - \sum_{i=1}^l y_i (\alpha_i^* - \alpha_i), \quad (4) \\ \text{s.t.} & \sum_{i=1}^l (\alpha_i^* - \alpha_i) = 0, \\ & 0 \leq \alpha_i^*, \quad \alpha_i \leq \frac{C}{l}, \end{aligned}$$

式中, $\alpha^{(*)} = (\alpha_1, \alpha_1^*, \alpha_2, \alpha_2^*, \dots, \alpha_l, \alpha_l^*)$ 即为 Lagrange 乘子. 该式是一个线性约束的二次规划问题,故存在全局最优解. 用序列最小优化 (sequential minimization optimization, SMO) 等多种方法可以求解出 $\alpha^{(*)}$, 则系数向量可以表示为

$$\mathbf{w} = \sum_{i=1}^l (\alpha_i^* - \alpha_i) \phi(\mathbf{x}_i), \quad (5)$$

然而, $\phi(\cdot)$ 并不好确定. 为了解决 $\phi(\cdot)$ 不好确定问题以及确定出来后特征空间的计算量太大导致的维度灾难问题, 考虑到(4)式中 $\phi(\cdot)$ 以内积的形式应用的, 因此引入核函数 $K(\mathbf{x}_i, \mathbf{x}_j)$, 有 $K(\mathbf{x}_i, \mathbf{x}_j) = (\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j))$ 成立. 根据 Hilbert-Schmidt 原理, 只要一种运算满足 Mercer 条件, 就可以作为核函数. 用核函数 $K(\mathbf{x}_i, \mathbf{x}_j)$ 来代替(4)式中的 $(\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j))$, 有

$$\begin{aligned} \min & \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l (\alpha_i^* - \alpha_i) (\alpha_j^* - \alpha_j) K(\mathbf{x}_i, \mathbf{x}_j) \\ & + \varepsilon \sum_{i=1}^l (\alpha_i^* + \alpha_i) - \sum_{i=1}^l y_i (\alpha_i^* - \alpha_i), \quad (6) \\ \text{s.t.} & \sum_{i=1}^l (\alpha_i^* - \alpha_i) = 0, \\ & 0 \leq \alpha_i^*, \quad \alpha_i \leq \frac{C}{l}. \end{aligned}$$

求得 $\alpha^{(*)} = (\alpha_1, \alpha_1^*, \alpha_2, \alpha_2^*, \dots, \alpha_l, \alpha_l^*)$ 后, 决策函

数为

$$f(\mathbf{x}) = \sum_{i=1}^l (\alpha_i^* - \alpha_i) K(\mathbf{x}_i, \mathbf{x}) + b, \quad (7)$$

其中 b 的值可利用 KKT(Karush-Kuhn-Tucker) 条件求得. 因为只有对应拉氏乘子满足式 $(\alpha_i^* - \alpha_i) \neq 0$ 的输入 \mathbf{x}_i 才对 $f(\mathbf{x})$ 有影响, 而且 $f(\mathbf{x})$ 也完全由这些输入向量决定, 所以这些向量被称为支持向量, 该算法被称为支持向量机.

常用的核函数有多项式核、高斯径向基核和 Sigmoid 核等. 本文中用到的高斯径向基核函数公式

$$K(x, x') = \exp(-\gamma \|x - x'\|^2), \quad (8)$$

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2p^2}\right). \quad (9)$$

3 支持向量机自由参数选取方法

应用 SVM 进行时间序列预测, 分为 3 步: 首先对原一维时间序列进行重构得到训练样本; 然后针对该训练样本进行 SVM 建模, 其中, 自由参数取值是否适当直接决定了所建 SVM 模型的优劣; 最后应用所得模型进行预测.

3.1 时间序列的重构

设时间序列 $\{x(t_1), x(t_2), \dots, x(t_i), \dots, x(t_l)\}$, 其中 $x(t_i)$ 表示初始时间后面第 i 个时间点的指标值. 时间序列预测基础是相空间的重构, 设嵌入维为 m , 步长 (或称延迟) 为 τ , 则重构后的输入输出对为

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]^T = \begin{bmatrix} x(t_1) & x(t_{1+\tau}) & \cdots & x(t_{1+(m-1)\tau}) \\ x(t_2) & x(t_{2+\tau}) & \cdots & x(t_{2+(m-1)\tau}) \\ \vdots & \vdots & \vdots & \vdots \\ x(t_N) & x(t_{N+\tau}) & \cdots & x(t_{N+(m-1)\tau}) \end{bmatrix}, \quad (10)$$

$$\mathbf{Y} = \begin{bmatrix} x(t_{1+m\tau}) \\ x(t_{2+m\tau}) \\ \vdots \\ x(t_{N+m\tau}) \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \quad (11)$$

其中, $N = l - m\tau$ 为重构后相空间点数. 建立时间序列的预测模型即是找出输入 \mathbf{X} 和输出 \mathbf{Y} 之间的函数 f 使 (1) 式成立. 延迟 τ 值和嵌入维 m 值的确定是关键问题, 目前已有多种成熟方法 [23].

3.2 基于残差信息冗余检验自由参数选取方法的提出

SVM 与神经网络相比能自动确定网络结构, 实现全局寻优; 但给自由参数设置不同的值将计算出不同的网络结构和模型. 因此自由参数的设置是否恰当直接决定了 SVM 模型优劣. 从 (3), (6) 式可知, 这些自由参数包括不敏感损失函数参数 ε 、惩罚参数 C , 以及核函数相关参数.

目前, 关于核函数的选择的共识是: 不同的问题适用的核函数不同; 在没有太多先验知识的情况下, 高斯径向基 (radial basis function, RBF) 核函数是首选, 因其对线性、非线性系统均有良好的逼近能力 [24].

在如何选择最优自由参数问题上, 目前主要研究有: Schölkopf 等提出了 ν -SVM 方法 [15], 通过设置 ν 直接控制误分点个数及支持向量个数实现 ε 的自动计算; Smola 和 Kwok 等 [16,17] 提出利用噪声方差来选取 ε ; 文献 [18] 提出利用输出值的宽度设置 C ; 文献 [19] 提出利用 K -折交叉验证法设置自由参数组; 文献 [20] 提出用留一法 (leave one out, LOO) 方法; 文献 [21] 利用统计学分布和置信区间概念改进了文献 [17, 18] 中 C 和 ε 的设置方法; 文献 [22] 提出支持向量机用于分类时高斯核函数参数核宽度 γ 的设置方法. 但这些方法中文献 [22] 只适用于支持向量机分类应用, 支持向量回归机中应用最多的方法是校验集方法, 该方法的效果依赖于校验集的选取, 其正确性缺乏理论基础; 另外所有这些方法只应用简单的残差相关统计量比如误差平方和 (mean squared error, MSE) 等; 目前为止没有文献考虑时间序列和一般 SVM 回归问题的不同特性. 针对这种情况, 本文提出了一种新的时间序列 SVM 自由参数选取方法——基于残差信息冗余检验的 SVM 自由参数选择方法.

时间序列和普通回归问题存在一个很大的不同: 普通回归问题中训练集的样本点通常假定是彼此独立的, 而时间序列训练集中的目标值间则存在着相关性; 这也是时间序列之所以能够利用历史数

据建立自回归模型并预测未来数据的原因. 因此, 一个有效的模型应该使得训练集残差不含冗余信息. 否则, 可以认为该模型没有很好的提取数据中的相关信息, 需要进一步优化.

时间序列本质上是一种特殊的动力学系统. Ljung 等 [25]、Zhang 等 [26]、以及 Mao 等 [27] 都研究了动力学系统辨识中应用残差相关性检验方法进行神经网络模型适用性检验, 解决了之前神经网络需要人为确定网络结构的问题, 实现了神经网络结构确定的科学化和自动化, 取得了很好效果. SVM 自由参数选取问题本质上是模型辨识 (model identification) 问题, 因为不同的自由参数值决定了所得的支持向量, 这些支持向量和所确定的核函数一起确定了最终的 SVM 模型. 但目前为止, 尚没有应用残差相关性检验方法来选取 SVM 自由参数的研究.

作者在移动网络业务量数据的 SVM 预测中提出了通过检验训练残差是否白噪声来选择自由参数方法, 并取得很好的应用效果 [28]. 但在将该方法推广到具有非线性特征数据如太阳黑子、Mackey-Glass 等时, 效果却不理想. 分析其原因在于: 通信网络语音业务量序列本质上是线性的, 因此残差线性无关时即表示残差是无关的; 但年均太阳黑子数等时间序列则是非线性的, 其残差线性不相关时可能还存在非线性的相关 [29]. 因此, 本文提出一种通过残差信息冗余检验来选取 SVM 自由参数方法, 接下来介绍应用全向相关函数 (ODCF) 检验残差信息冗余及具体检验的方法, 并从例子和理论推导两方面给出了该信息冗余检验方法有效性的证明.

3.3 残差信息冗余检验统计量的选取

信息论中的互信息、冗余、广义冗余等统计量都是衡量随机变量相关性的 [30], 但这些指标没有明确的分布规律, 因此无法用来进行某数据有无信息冗余的假设检验. 目前, 普遍采用的在关联积分概念基础上发展而来的检验冗余的 BDS 统计量存在两个问题 [31]: 首先, 应用该 BDS 统计量要求训练集容量要大, 一般要求大于 1000. 其次, BDS 统计量只能用于检验自相关性, 无法检验两个随机变量间的相关. Zhang, Zhu 和 Longden 等 [32,33] 提出的基于 ACF, CCF 的 ODCF 系列统计量, 可以检验包括线性和非线性关系各种相关, 而且适用于包

括小容量的任意容量的模型检验中.

ODCF 统计量包括 ODCCF (omni-directional cross correlation function) 和 ODACF (omni-directional auto-correlation function), 而 ODACF 又可以看作 ODCCF 的一个特例, 即两个随机序列同时情况. 下面是 ODCCF 的计算方法.

给定两个随机序列 $a(t), b(t)$, 令

$$\alpha(t) = |a'(t)| = \left| a(t) - \frac{1}{N} \sum_{t=1}^N a(t) \right|, \quad (12)$$

$$\beta(t) = |b'(t)| = \left| b(t) - \frac{1}{N} \sum_{t=1}^N b(t) \right|,$$

则 ODCCF 有以下四种相关系数计算函数:

$$r_{\alpha\beta}(\tau) = \frac{\sum_{t=\tau+1}^N (\alpha(t-\tau)\beta(t))}{\sqrt{\sum_{t=1}^N (\alpha(t))^2 \sum_{t=1}^N (\beta(t))^2}}, \quad (13)$$

$$r_{\alpha b'}(\tau) = \frac{\sum_{t=\tau+1}^N (\alpha(t-\tau)b'(t))}{\sqrt{\sum_{t=1}^N (\alpha(t))^2 \sum_{t=1}^N (b'(t))^2}}, \quad (14)$$

$$r_{a'b'}(\tau) = \frac{\sum_{t=\tau+1}^N (a'(t-\tau)b'(t))}{\sqrt{\sum_{t=1}^N (a'(t))^2 \sum_{t=1}^N (b'(t))^2}}, \quad (15)$$

$$r_{a'\beta}(\tau) = \frac{\sum_{t=\tau+1}^N (a'(t-\tau)\beta(t))}{\sqrt{\sum_{t=1}^N (a'(t))^2 \sum_{t=1}^N (\beta(t))^2}}. \quad (16)$$

综合以上 4 种相关函数, 得到一个用来衡量相关性的 ODCCF 统计量函数 $\rho_{ab}(\tau)$:

$$|\max(r_{\alpha\beta}(\tau), r_{\alpha b'}(\tau), r_{a'b'}(\tau), r_{a'\beta}(\tau))|$$

$$> |\min(r_{\alpha\beta}(\tau), r_{\alpha b'}(\tau), r_{a'b'}(\tau), r_{a'\beta}(\tau))|,$$

则

$$\rho_{ab}(\tau) = \max(r_{\alpha\beta}(\tau), r_{\alpha b'}(\tau), r_{a'b'}(\tau), r_{a'\beta}(\tau)), \quad (17)$$

否则

$$\rho_{ab}(\tau) = \min(r_{\alpha\beta}(\tau), r_{\alpha b'}(\tau), r_{a'b'}(\tau), r_{a'\beta}(\tau)). \quad (18)$$

当 $a(t) = b(t)$ 时, 上式称为 ODACF.

根据中心极限定理及文献 [26], 当残差不含冗余信息时, 应有

$$\rho_{\varepsilon\varepsilon}(\tau) = 1, \quad \tau = 0, \quad (19)$$

$$\rho_{\varepsilon\varepsilon}(\tau) \sim N\left(0, \frac{1}{n}\right), \quad \text{其他};$$

$$\rho_{y\varepsilon}(\tau) \neq 0, \quad \tau = 0, \quad (20)$$

$$\rho_{y\varepsilon}(\tau) \sim N\left(0, \frac{1}{n}\right), \quad \text{其他}.$$

3.4 基于残差信息冗余检验的自由参数选取方法及证明

时间序列模型作为一种没有输入信息的动力学系统可用下面数学公式表示

$$y(t) = f(\mathbf{y}^{t-1}) + e(t)$$

$$= f(y(t-1), \dots, y(t-t_y)) + e(t). \quad (21)$$

在进行模型辨识时, 得到真实模型的估计模型 $\hat{f}(\cdot)$, 并应用该估计模型得到输出的估计值

$$\hat{y}(t) = \hat{f}(\mathbf{y}^{t-1}), \quad \varepsilon(t) = y(t) - \hat{y}(t), \quad (22)$$

其中, $\varepsilon(t)$ 是在第 t 个时间点上的预测误差或称残差. 对于一个动力学系统, 一个有效的模型应该把真实系统中的有效信息全部提取出来, 剩余的残差应该是白噪声序列 [28]. 反之, 如果所得残差与自身或者输出有相关, 则表示该训练模型不是最优的, 应该进一步训练, 此时的残差可以表示为

$$\varepsilon(t) = \hat{g}(\mathbf{y}^{t-1}, \mathbf{e}^{t-1}) + e(t). \quad (23)$$

$\hat{g}(\cdot)$ 可能是线性或非线性的, 它应该通过进一步训练提取到最终的估计模型 $\hat{f}(\cdot)$ 中. 由 (23) 式, 可以做如下推导:

$$\varepsilon(t) = \hat{g}(\dots, y(t-p), \dots) + e(t)$$

$$= \hat{g}(\dots, \hat{f}(y(t-i-p)) + \varepsilon(t-p), \dots) + e(t). \quad (24)$$

由该式可以看出, 若输出和残差相关, 则残差将表现出自相关. 但这只是理论上理想情况, 实际上, 如果函数 $g(t)$ 的方差远小于噪声 $e(t)$ 的方差, 则残差的自相关可能检验不出来, 本节接下来将从理论分析和例子两方面给出这种情况的证明. 因此, 论文提出, 在检验残差是否含信息冗余时, 需要同时检验残差自身相关性和残差与输出的相关性才能最大程度保证模型有效性和自由参数最优性.

定理 令 $\text{var}(x)$ 代表随机变量 x 的方差, 存在数值 k , 当 $\frac{\text{var}(e)}{\text{var}(g)} > k$ 时, $\rho_{\varepsilon\varepsilon}(\tau)$ 总落在给定置信度下置信区间内.

证明 根据文献 [32] 中的 Definition 3, $\rho_{\varepsilon\varepsilon}(\tau)$ 值是 4 个 ODCCF 值中的绝对值最大的值, 即

$$|\rho_{\varepsilon\varepsilon}(\tau)| = \max(|r_{|e'|e'}(\tau)|, |r_{|e'|g'}(\tau)|, |r_{e'e'}(\tau)|, |r_{e'|g'}(\tau)|). \quad (25)$$

因此下面从四种相关系数分别证明:

$$\begin{aligned} & r_{|e'|e'}(\tau) \\ &= \frac{\sum_{t=\tau+1}^N |g'(t-\tau) + e'(t-\tau)||g'(t) + e'(t)|}{\sqrt{\text{var}(|g' + e'|)\text{var}(|g' + e'|)}} \\ &\leq \frac{\sum_{t=\tau+1}^N (|g'(t-\tau)| + |e'(t-\tau)|)(|g'(t)| + |e'(t)|)}{\sqrt{\text{var}(|g' + e'|)\text{var}(|g' + e'|)}} \\ &\approx \frac{\sum_{t=\tau+1}^N |g'(t-\tau)||g'(t)| + \sum_{t=\tau+1}^N |e'(t-\tau)||g'(t)|}{\sqrt{\text{var}(|g' + e'|)\text{var}(|g' + e'|)}} \\ &= \frac{r_{|g'|g'}(\tau)}{\frac{\text{var}(|g' + e'|)}{\text{var}(|g'|)}} + \frac{r_{|e'|g'}(\tau)}{\sqrt{\frac{\text{var}(|g' + e'|)\text{var}(|g' + e'|)}{\text{var}(|g'|)\text{var}(|e'|)}}} \\ &\approx \frac{r_{|g'|g'}(\tau)}{\frac{\text{var}(|e'|)}{\text{var}(|g'|)}} + \frac{r_{|e'|g'}(\tau)}{\sqrt{\frac{\text{var}(|e'|)}{\text{var}(|g'|)}}}, \end{aligned} \quad (26)$$

$$\begin{aligned} & r_{|e'|g'}(\tau) \\ &= \frac{\sum_{t=\tau+1}^N |g'(t-\tau) + e'(t-\tau)|(g'(t) + e'(t))}{\sqrt{\text{var}(|g' + e'|)\text{var}(g' + e')}} \\ &\approx \frac{\sum_{t=\tau+1}^N (|e'(t-\tau)|)(g'(t) + e'(t))}{\sqrt{\text{var}(|g' + e'|)\text{var}(g' + e')}} \\ &\approx \frac{\sum_{t=\tau+1}^N |e'(t-\tau)|g'(t)}{\sqrt{\text{var}(|g' + e'|)\text{var}(g' + e')}} \\ &= \frac{r_{|e'|g'}(\tau)}{\sqrt{\frac{\text{var}(|g' + e'|)\text{var}(g' + e')}{\text{var}(|e'|)\text{var}(g')}}} \\ &\approx \frac{r_{|e'|g'}(\tau)}{\sqrt{\frac{\text{var}(e')}{\text{var}(g')}}}, \end{aligned} \quad (27)$$

$$r_{e'e'}(\tau)$$

$$\begin{aligned} &= \frac{\sum_{t=\tau+1}^N (g'(t-\tau) + e'(t-\tau))(g'(t) + e'(t))}{\sqrt{\text{var}(g' + e')\text{var}(g' + e')}} \\ &\approx \frac{\sum_{t=\tau+1}^N g'(t-\tau)g'(t) + \sum_{t=\tau+1}^N e'(t-\tau)g'(t)}{\sqrt{\text{var}(g' + e')\text{var}(g' + e')}} \\ &= \frac{r_{g'g'}(\tau)}{\frac{\text{var}(g' + e')}{\text{var}(g')}} + \frac{r_{e'g'}(\tau)}{\sqrt{\frac{\text{var}(g' + e')\text{var}(g' + e')}{\text{var}(g')\text{var}(e')}}} \\ &\approx \frac{r_{g'g'}(\tau)}{\frac{\text{var}(e')}{\text{var}(g')}} + \frac{r_{e'g'}(\tau)}{\sqrt{\frac{\text{var}(e')}{\text{var}(g')}}}, \end{aligned} \quad (28)$$

$$\begin{aligned} & r_{\varepsilon'|e'}(\tau) \\ &= \frac{\sum_{t=\tau+1}^N (g'(t-\tau) + e'(t-\tau))|g'(t) + e'(t)|}{\sqrt{\text{var}(g' + e')\text{var}(|g' + e'|)}} \\ &\approx \frac{\sum_{t=\tau+1}^N (g'(t-\tau) + e'(t-\tau))(|g'(t)| + |e'(t)|)}{\sqrt{\text{var}(g' + e')\text{var}(|g' + e'|)}} \\ &\approx \frac{\sum_{t=\tau+1}^N g'(t-\tau)|g'(t)| + \sum_{t=\tau+1}^N e'(t-\tau)|g'(t)|}{\sqrt{\text{var}(g' + e')\text{var}(|g' + e'|)}} \\ &= \frac{r_{g'|g'}(\tau)}{\sqrt{\frac{\text{var}(g' + e')\text{var}(|g' + e'|)}{\text{var}(|g'|)\text{var}(g')}}} \\ &\quad + \frac{r_{e'|g'}(\tau)}{\sqrt{\frac{\text{var}(g' + e')\text{var}(|g' + e'|)}{\text{var}(|g'|)\text{var}(e')}}} \\ &\approx \frac{r_{g'|g'}(\tau)}{\sqrt{\frac{\text{var}(e')\text{var}(|e'|)}{\text{var}(|g'|)\text{var}(g')}}} + \frac{r_{e'|g'}(\tau)}{\sqrt{\frac{\text{var}(|e'|)}{\text{var}(|g'|)}}}. \end{aligned} \quad (29)$$

设 $\frac{\text{var}(e')}{\text{var}(g')} = k$.

1) 对于 (27) 式和 (28) 式, 分子均是相关系数, 满足 $|r_{|e'|g'}(\tau)| \leq 1, |r_{e'g'}(\tau)| \leq 1, |r_{g'g'}(\tau)| \leq 1, |r_{e'e'}(\tau)| \leq 1$, 分母为 $\sqrt{\frac{\text{var}(e')}{\text{var}(g')}} = \sqrt{k}$ 和 $\frac{\text{var}(e')}{\text{var}(g')} = k$, 当比值 k 大于某一值时, 将导致 $r_{|e'|e'}(\tau), r_{e'e'}(\tau)$ 落在置信区间内.

2) 对于前面 (26) 式和 (29) 式, 分子也都是相关系数, 绝对值小于 1, 分母含 $\frac{\text{var}(|e'|)}{\text{var}(|g'|)}$ 项, 可以证明, 若随机变量 $x \sim N(0, \sigma^2)$, 则其绝对值的均

值为 $\frac{2\sigma}{\sqrt{2\pi}}$, 方差为 $\frac{(\pi-2)\sigma^2}{\pi}$ (证明见附录), 因此, 有 $\frac{\text{var}(|e'|)}{\text{var}(|g'|)} = \frac{\text{var}(e')}{\text{var}(g')} = k$. 当比值 k 大于某一值时, 将导致 $r_{|e'|}|e'|(\tau), r_{e'|e'|}(\tau)$ 落在给定置信度下的置信区间内.

3) 由 $\rho_{\varepsilon\varepsilon}(\tau)$ 的绝对值是上面 4 个系数绝对值最大值, 综合 1), 2) 知, 这样的 k 存在.

下面举例说明. 设有如下非线性模型:

$$y(t) = 0.8y(t-1) + 0.8y^2(t-1) + e(t), \quad (30)$$

其中, $e(t)$ 为一符合高斯分布的随机量, 有 $e(t) \sim N(0, 0.001)$.

设有如下 4 个残差公式:

$$\begin{aligned} \varepsilon_1(t) &= e(t), \\ \varepsilon_2(t) &= 0.6e(t-1) + e(t), \\ \varepsilon_3(t) &= 0.1y(t-1) + e(t), \\ \varepsilon_4(t) &= 0.02y(t-1) + e(t). \end{aligned} \quad (31)$$

计算以上 4 种残差下的全向相关函数, 包括全向互相关和全向自相关函数如图 1—图 4 所示, 其

中的虚线表示 2 倍标准差即置信度采用 95% 时的置信区间. 图 1 显示, 模型残差为噪声时, 各滞后阶下全向相关函数值都落在置信区间内, 表示此模型残差不含冗余信息, 模型通过有效性检验, 这与实际情况一致. 而图 2—图 4 中, 都有至少 1 个相关函数值落在置信区间外. 图 2 显示, 在滞后阶 $\text{lag} = 2$ 时, 残差全向自相关函数 ODACF 值 $\rho_{\varepsilon\varepsilon}(2)$ 落在置信区间外, 表示残差含有冗余信息, 这与实际情况相符. 图 3 对应的第 3 种残差函数 $g(\cdot)$ 是输出过去值的函数, 在 $\text{lag} = 2, \rho_{\varepsilon\varepsilon}(2)$ 落在 2 倍标准差外, 当然, 其 $\rho_{y\varepsilon}(2)$ 也落在 2 倍标准差外. 对于 $\varepsilon_4(t), g(t) = 0.02y(t-1)$ 仍然是关于输出过去值的函数, 与 $\varepsilon_3(t)$ 的唯一区别是系数由 0.1 减小为 0.02, 但图 4 显示, ODACF $\rho_{\varepsilon\varepsilon}(\tau)$ 已经全部落在置信区间内, 分析原因就在于在第 4 种残差中, $\text{var}(e) \gg \text{var}(g)$, 有 $\frac{\text{var}(e)}{\text{var}(g)} = 781$. 但对于 $\varepsilon_4(t), \rho_{y\varepsilon}(12), \rho_{y\varepsilon}(13)$ 仍落在置信区间外. 因此, 检验残差是否含信息冗余需要同时检验自相关函数 $\rho_{\varepsilon\varepsilon}(\tau)$ 和互相关函数 $\rho_{y\varepsilon}(\tau)$, 才能最好的检验残差信息冗余情况, 从而得到最优的自由参数和模型.

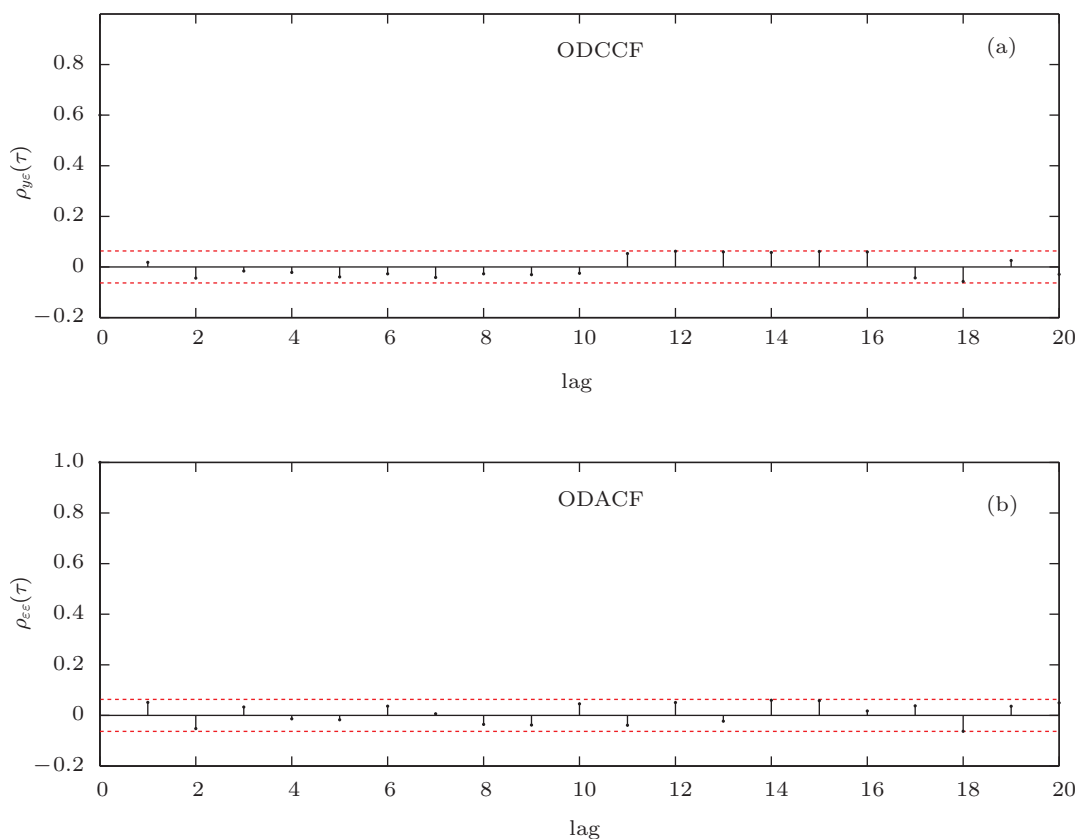


图 1 ε_1 的有效性验证 (a) $\rho_{y\varepsilon}(\tau)$; (b) $\rho_{\varepsilon\varepsilon}(\tau)$

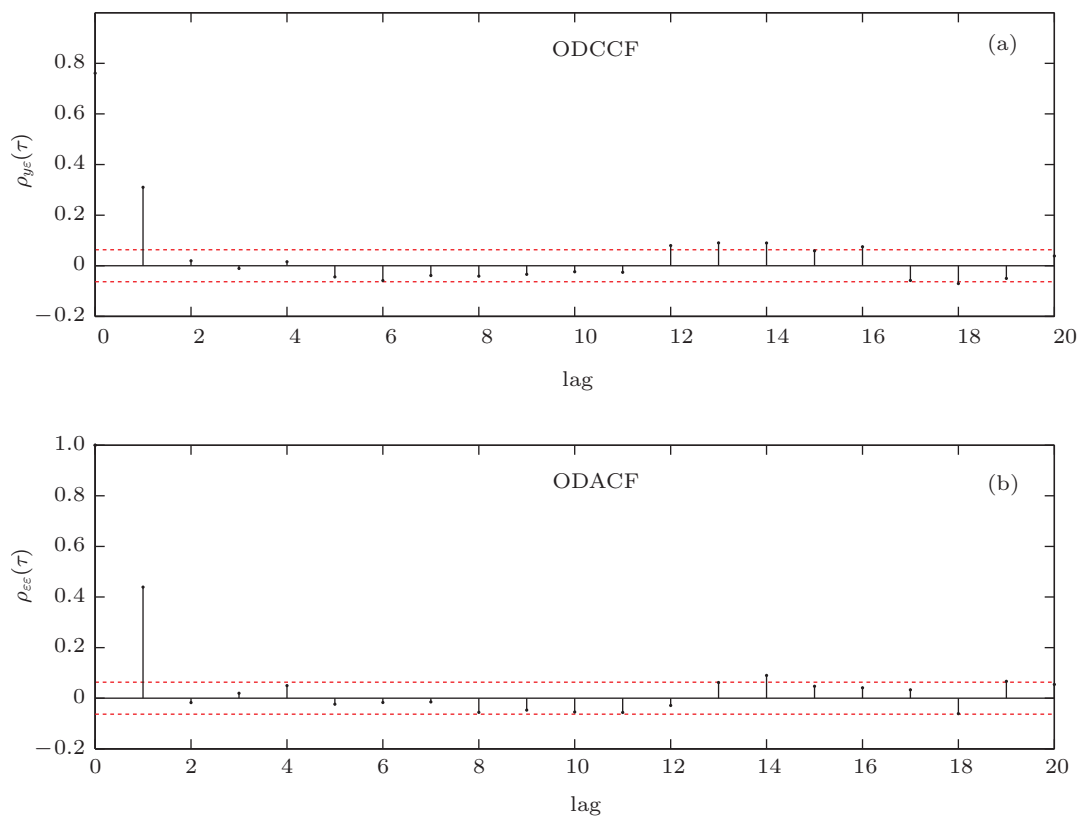


图2 ε_2 的有效性验证 (a) $\rho_{y\varepsilon}(\tau)$; (b) $\rho_{\varepsilon\varepsilon}(\tau)$

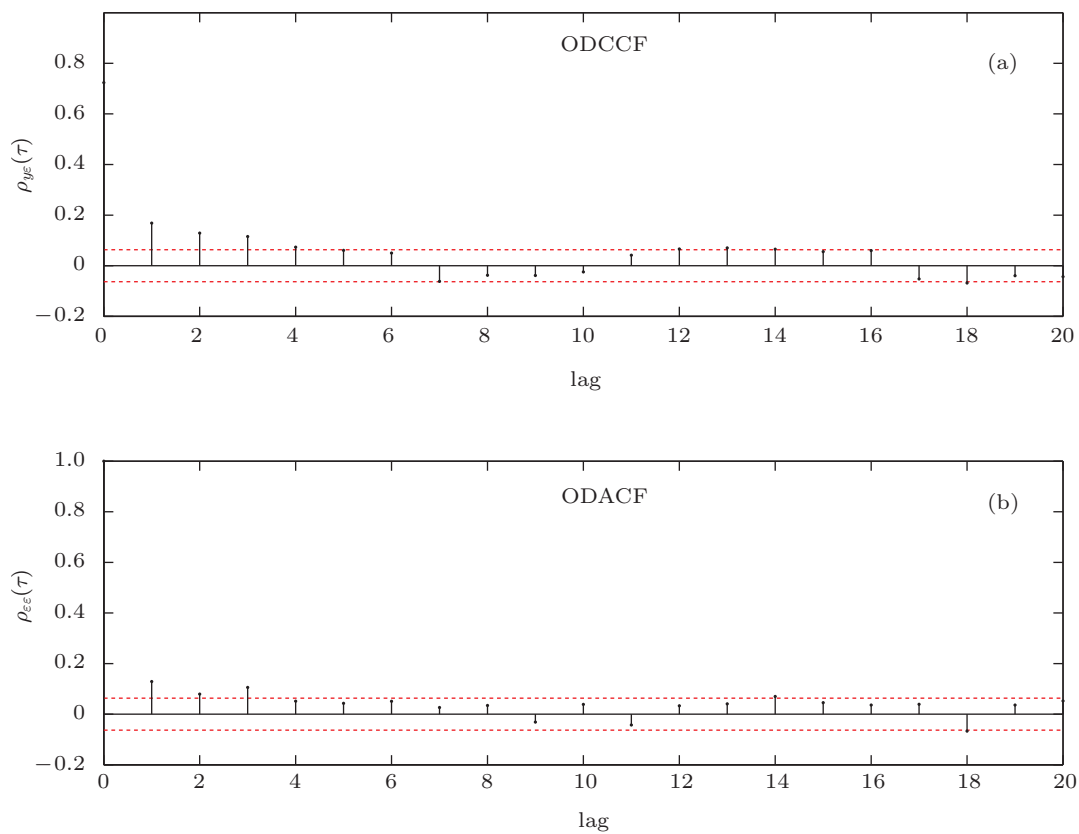


图3 ε_3 的有效性验证 (a) $\rho_{y\varepsilon}(\tau)$; (b) $\rho_{\varepsilon\varepsilon}(\tau)$

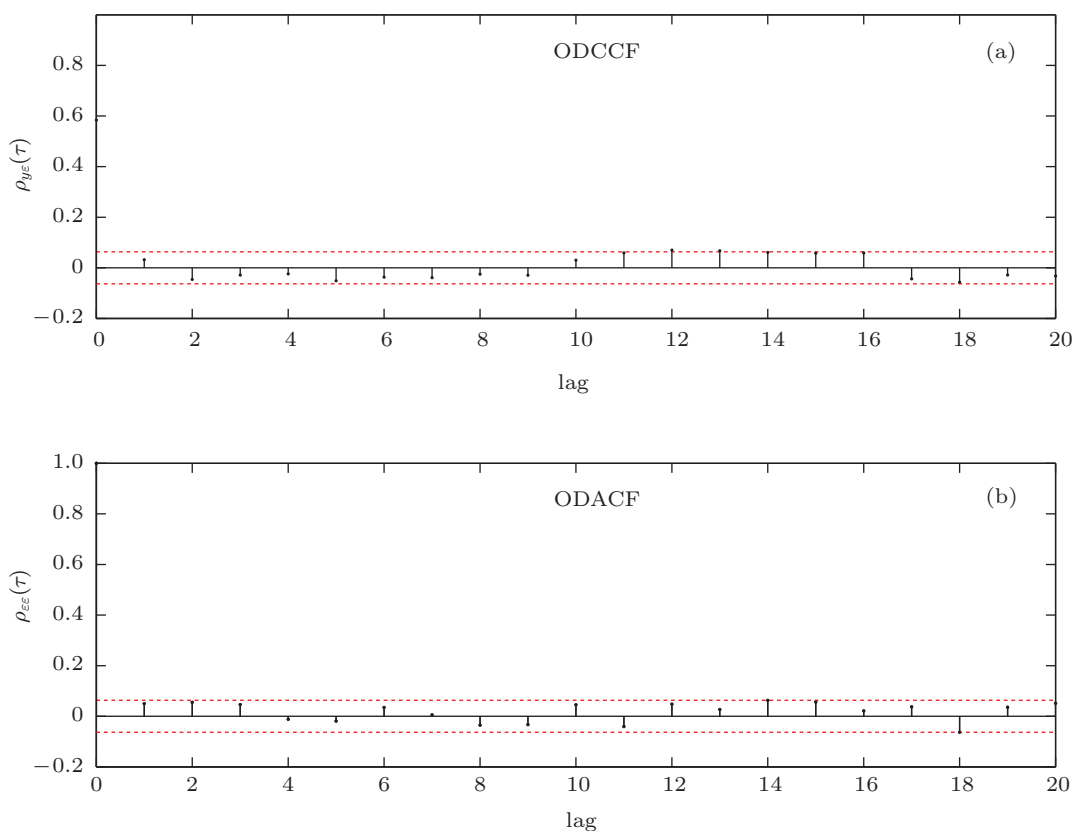


图4 ε_4 的有效性验证 (a) $\rho_{y\varepsilon}(\tau)$; (b) $\rho_{\varepsilon\varepsilon}(\tau)$

4 实验和结果分析

本节首先描述通过应用 ODCF 检验残差信息冗余来选取最优 SVM 自由参数的实施过程, 然后介绍在太阳黑子序列和 Mackey-Glass 序列上实验过程和结果.

支持向量机采用 RBF 核时, 核函数宽度相关参数 γ 和惩罚参数 C 对建模结果影响很大, 而不敏感损失区域宽度 ε 的值则对建模结果影响不显著^[21]. 考虑到 ε 的值与支持向量的个数相关, 本文采用 ν -SVR^[15], 该类 SVM 通过设 ν 的值可自动计算得到 ε 的值, 不妨设 $\nu = 0.1$ ^[21]. 对于 γ 的设置, 应用 Cherkassky 等^[21] 提出的方法: 高斯核宽度应该反映训练数据的输入范围, 且为输入数据的 0.2 到 0.5 倍时是最优的; 对于多维输入向量, 对于归一化到 $[0, 1]$ 的输入数据有

$$p^m \sim 0.2-0.5, \quad \text{其中 } m \text{ 为维数.} \quad (32)$$

应用该方法的过程如下: 令 $\nu = 0.1$, 采用 (32) 式给出计算方法得到 (9) 式中核函数宽度参数 p 值并换

算成 (8) 式 γ 的值, 该参数值确定了函数的复杂度和模型的拟合能力. 然后改变参数 C 的值, 直到模型通过在某个较小置信区间上训练残差无冗余信息的假设检验, 此时自由参数为最优的, 所得模型为最佳 SVM 模型.

4.1 年均太阳黑子数预测

年均太阳黑子数是一个典型的非线性动力学系统产生的非线性序列, 因此常被用来检验和比较不同的学习方法及其预测效果. 为了使预测结果具有更直观的可比性, 本文中太阳黑子预测实验采用与文献 [34—36] 相同的实验设置, 所用数据来自 SIDC(solar influences data analysis center). 软件采用 matlab7.1 和 libsvm2.83^[37], 且将 libsvm2.83 嵌入到 matlab7.1 中实现循环嵌套控制. 实验过程如下:

1) 从原一维时间序列中重构相空间. 很多研究表明年均太阳黑子时间序列存在混沌特性, 计算其奇异吸引子关联维数 $D < 2$, 应用 Takens 定

理 $m \geq 2D + 1$, 论文采用嵌入维 $m = 5$, 重构年均太阳黑子序列相空间.

2) 应用基于残差信息冗余检验方法选取最优自由参数.

i. 设 v -SVM 中支持向量调节参数 $v = 0.1$.

ii. 设置外层循环变量即指示核函数宽度的参数值 γ . 维度 $m = 5$ 时, 应用 (32) 式计算得高斯核表达式 (9) 中参数 p 的范围, 并换算成表达式 (8) 中的 γ 的取值范围: 0.5—0.8. 以初值为 0.5 步长为 0.1 取该区间内 γ 的各值.

iii. 设外层循环变量值 $\gamma = 0.5$, 设置内层循环初始值 $C = 1$. 以步长 1 对 C 进行循环, 发现在 2 倍标准差下没有使得两种 ODCF 量 $\rho_{y\epsilon}(\tau)$ 和 $\rho_{\epsilon\epsilon}(\tau)$ 都通过残差无信息冗余检验的 C . 而在 3 倍标准差下有大量通过检验 C . 因此, 循环减小置信区间的大小, 步长采用 0.05 倍标准差, 直到没有 C 使得 ODCF 量 $\rho_{y\epsilon}(\tau)$ 和 $\rho_{\epsilon\epsilon}(\tau)$ 通过无关性检验. 发现置信区间在 2.65 倍标准差时是通过无关性检验的最小置信区间, 而此时通过检验的第一个 C 值为 $C = 1589$ 时, 可以认为此模型是 $\gamma = 0.5$ 时最优模型.

iv. 改变 γ 值并重复以上 iii 中的过程. $\gamma = 0.6$ 时, 通过无关检验的最小置信区间是 2.85 倍标准差. $\gamma = 0.7$ 时通过无关检验的最小置信区间是 2.8 倍标准差, $\gamma = 0.8$ 是 2.85 倍标准差. 因此最佳模型对应的自由参数组为 $\gamma = 0.5, C = 1589, v = 0.1$.

3) 应用所得最佳 SVM 模型对 1956—1979 年的太阳黑子数进行预测. 结果如图 5 所示. 对该预测性能进行评价, 首先采用文献 [34—36] 中采用的 NMSE 指标, 该指标的计算方法如下:

$$NMSE = \frac{1}{\delta^2 l} \left(\sum_{i=1}^l (y_i - \hat{y}_i)^2 \right), \quad (33)$$

$$\delta^2 = \frac{1}{l-1} \left(\sum_{i=1}^l (y_i - \bar{y})^2 \right).$$

预测性能及其与其他文献中预测方法的比较如表 1 所示, 可以看出, 以 NMSE 作为预测性能评价指标, 本论文方法预测精度远高于文献 [35,36] 中所提方法, 也优于文献 [34] 中的方法. 统计量 NMSE 说明的是预测结果与真实值的总体偏移情况, 为了进一步说明误差在各点上的分布情况, 再计算 Pearson

相关系数 ρ , 其计算公式如下:

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{D_X} \sqrt{D_Y}} = \frac{E\{(X - E(X))(Y - E(Y))\}}{\sqrt{D_X} \sqrt{D_Y}}, \quad (34)$$

其中

$$E(X) = \frac{1}{n} \sum_{i=1}^n X_i,$$

$$D(X) = \frac{1}{n-1} \sum_{i=1}^n (X_i - E(X))^2.$$

计算得到 1956—1979 年太阳黑子预测值与实际值的相关系数 $\rho = 0.93$, 接近最大值 1, 说明实际值与预测值有一致的分布和走向, 预测误差在各点上的分布是均匀的, 该特性从图 5 可以看出.

表 1 1956—1979 年年均太阳黑子数预测性能

Methods	NMSE/%
基于残差独立检验的 SVM	13.5
Benchmark ^[34]	15.4
Benchmark ^[35]	35.0
Benchmark ^[36]	28.0

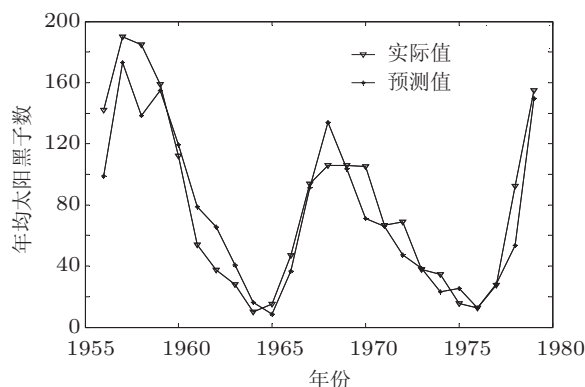


图 5 1956—1979 年年均太阳黑子数实际值和预测值

4.2 Mackey-Glass 序列预测

第 2 个应用是典型的高维混沌系统 — Mackey-Glass 时延差分公式产生的序列.

$$\frac{dx(t)}{dt} = -0.1x(t) + \frac{0.2x(t - t_d)}{1 + x(t - t_d)^{10}}. \quad (35)$$

该式是建模血细胞生成模型时采用的, 已经证明, 当 $t_d \geq 17$ 时所产生的序列为高维混沌系统. 该序列也是评价预测模型的经典基准数据 [12,38], 图 6

为 $t_d = 30$ 时用四阶龙格 - 库塔法 (Runge-Kutta method) 产生的前 1200 个数据点.

1) 采用与文献 [38] 相同的相空间重构方法: 嵌入维 $m = 6$ 和步长 $\tau = 6$, 采用一步预测, 即将输入 \mathbf{X}_N 代入模型得到 $x(t_{N+m\tau})$ 的预测值. 把整个数据集分为 6 个独立的集合, 第 1 个集合 S_1 包含点 $x_{1+(d-1)\tau}$, $d = 1, 2, \dots$, 第 2 个集合 S_2 包含 $x_{2+(d-1)\tau}$, 依此类推, 第 6 个集合 S_6 包含点 $x_{d\tau}$. 文献 [41] 将 S_1 的前 100 个点用作训练集, S_2 的前 100 个点用作校验集, 将使得校验集误差平方和最小的自由参数值作为最优.

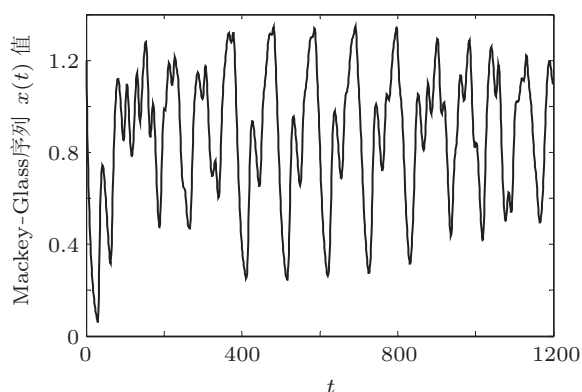


图 6 Mackey-Glass 时间序列 MG17 的前 1200 个点

2) 应用基于信息冗余检验方法选择 SVM 最优自由参数.

- i. 设 ν -SVM 中支持向量调节参数 $\nu = 0.1$.
- ii. 维度 $m = 6$ 时, 应用 (32) 式计算高斯核表达式 (9) 中核宽度参数 p 的范围, 并换算成表达式 (8) 中的 γ 的取值范围: 0.5—0.7, 以步长 0.1 取该区间的各值.
- iii. 设外层循环变量值 $\gamma = 0.5$, 设内层循环参数初始值 $C = 1$. 对 C 进行循环, 检验每次循环所得 SVM 模型拟合残差是否无信息冗余. 置信区间为 2 倍标准差下第一个残差无信息冗余的 C 值 32.
- iv. 在 $\gamma = 0.6, \gamma = 0.7$ 时重复这个过程, 得到 2 倍标准差下的的一个无信息冗余的 C 分别是 36, 18. 说明在 3 个 $\gamma = 0.5$ 值下, 都有合适的模型. 不妨取 γ 最小值下的模型.

3) 设置自由参数值为: $\gamma = 0.5, \nu = 0.1, C = 32$, 解优化问题得到最优模型. 应用所得模型对 S_1 第 101—200 个点进行预测. 所得预测结果如图 7 所示.

对该预测性能进行评价, 首先采用文献 [38] 采用的 RMSE(均方根误差), 该指标的计算方法为

$$RMSE = \sqrt{\frac{1}{l} \left(\sum_{i=1}^l (y_i - \hat{y}_i)^2 \right)}. \quad (36)$$

该方法及校验集方法在测试集上预测值 RMSE 如表 2 所示. 表 2 中同时列出了蔡俊伟等学者在文献 [5] 及崔万照等学者在文献 [6] 中 Mackey-Glass 实验预测性能.

表 2 Mackey - Glass 序列预测性能评价

方法	RMSE
基于残差独立检验 SVM	0.016
用 S_2 作校验集	0.018
基于选择性支持向量机集成 [5]	0.04(MSE = 0.0016)
参数未特别优化 [6]	0.058(MSE = 0.0034)

由表 2 可以看出, 基于残差信息冗余检验方法预测精度高于校验集方法, 并优于文献 [5,6] 中的方法. RMSE 说明了预测结果与真实值的总体偏移情况, 为了进一步说明误差在各点上的分布情况, 再计算 Pearson 相关系数 $\rho = 0.998$, 接近最大值 1, 说明实际值与预测值有一致的分布和走向, 预测误差在各点上的分布是均匀的, 这个计算结果和图 6 所示一致.

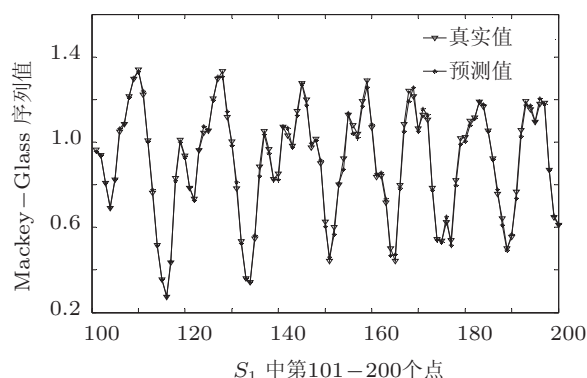


图 7 Mackey-Glass S_1 序列第 101—200 个点的预测值与实际值

分析本文提出自由参数选择方法预测结果好于校验集方法的原因, 从图 8 可以看出: 随着 C 的增大, 测试集误差 (此处采用统计量均方误差 mean squared error, MSE) 经历了先减小后增大的变化过程, 并在 $C = 32$ 时达到最小值; 这说明随着 C 的增大模型出现了先欠拟合到最佳拟合再到过拟合的变化, 并在 $C = 32$ 附近达到最佳拟合模型. 但校

验集上的 MSE 在 $C = 32$ 时仍呈快速下降趋势, 直到 $C = 78$ 以后才收敛. 因此, 校验集 MSE 的变化和测试集 MSE 的变化并不同步: 在测试集 MSE 最小时, 校验集还处在明显下降阶段, 而校验集 MSE 最小时实际上已经处于过拟合状态, 用校验集方法来选取最佳自由参数的方法存在不合理之处. 本文提出的基于残差信息冗余检验的自由参数选择方法则解决了这个问题: 第一个使得残差不含信息冗余的 C 值 $C = 32$, 此时正是测试集误差 MSE 最小值.

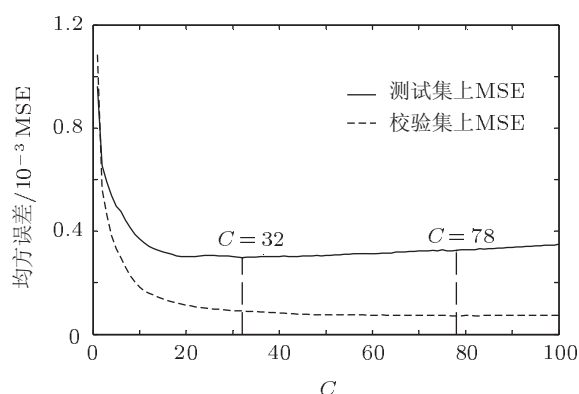


图 8 参数 $C = 1-100$ 时测试集和校验集上的 MSE 变化情况

5 结论

时间序列预测广泛应用于很多领域. 基于结构风险最小原则的支持向量机技术因其完备的理论基础, 解决了传统神经网络存在的固有问题, 越来越多的应用在时间序列预测中. 但支持向量机应用的一个关键和难点问题是自由参数如何选取问题: 参数选取是否适当直接决定了训练模型优劣, 不恰当的自由参数设置同样会导致欠拟合或过拟合. 常用的校验集方法计算量太大外, 其正确性缺乏理论

依据. 论文提出了通过检验残差信息冗余选取自由参数的方法, 并提出采用 ODCF 来检验残差信息冗余; 给出了应用 ODCF 检验残差信息冗余的方法, 并从理论分析和数值仿真两方面给出证明. 在两个经典非线性时间序列上进行了实验, 实验结果表明该方法可取得优于校验集方法及文献记载的方法的预测结果.

附录

设某随机变量 x 符合正态分布, 且均值为 0, 即有

$$x \sim N(0, \sigma^2),$$

则下面给出 $|x|$ 的均值和标准差推导过程:

$$\begin{aligned} E(|x|) &= \int_{-\infty}^{+\infty} |x| \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} dx \\ &= 2 \int_0^{+\infty} \frac{x}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} dx \\ &= \frac{-2\sigma}{\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \Big|_0^{+\infty} \\ &= \frac{2\sigma}{\sqrt{2\pi}}, \end{aligned} \tag{A1}$$

$$\begin{aligned} D(|x|) &= \int_{-\infty}^{+\infty} \left(|x| - \frac{2\sigma}{\sqrt{2\pi}} \right)^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} dx \\ &= \int_{-\infty}^{+\infty} |x|^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} dx \\ &\quad - \int_{-\infty}^{+\infty} \frac{4|x|}{2\pi} e^{-\frac{x^2}{2\sigma^2}} dx \\ &\quad + \int_{-\infty}^{+\infty} \frac{2\sigma^2}{\pi} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} dx \\ &= \sigma^2 + \frac{2\sigma^2}{\pi} - \int_0^{+\infty} \frac{4x}{\pi} e^{-\frac{x^2}{2\sigma^2}} dx \\ &= \sigma^2 + \frac{2\sigma^2}{\pi} + \frac{4\sigma^2 e^{-\frac{x^2}{2\sigma^2}}}{\pi} \Big|_0^{+\infty} \\ &= \frac{(\pi - 2)\sigma^2}{\pi}. \end{aligned} \tag{A2}$$

[1] Box G E P, Jenkins G M, Reinsel G C 2005 *Time Series Analysis: Forecasting and control*. (Posts & Telecom Press)
 [2] Yang H, Wang R 2011 *Acta Phys. Sin.* **60** 070508 (in Chinese) [杨红, 王瑞 2011 物理学报 **60** 070508]
 [3] Chen Q, Ren X M 2010 *Acta Phys. Sin.* **59** 2310 (in Chinese) [陈强, 任雪梅 2010 物理学报 **59** 2310]
 [4] Shi Z W; Han M 2007 *IEEE Trans. Neural Netw.* **18** 359
 [5] Cai J W, Hu S S, Tao H F 2007 *Acta Phys. Sin.* **56** 6820 (in Chinese) [蔡俊伟, 胡寿松, 陶洪峰 2007 物理学报 **56** 6820]
 [6] Cui W Z, Zhu C C, Bao W X, Liu J H 2005 *Acta Phys. Sin.* **54**

3009 (in Chinese) [崔万照, 朱长纯, 保文星, 刘君华 2005 物理学报 **54** 3009]
 [7] Ye M Y, Wang X D 2004 *Chin. Phys.* **13** 454
 [8] Vapnik V N 1999 *The Nature of Statistical Learning Theory* (2nd Ed.) (New York: Springer)
 [9] Vapnik V N 1998 *Statistical Learning Theory* (New York: Wiley)
 [10] Sapankevych N I, Sankar R 2009 *IEEE Comput. Intell. M.* **5** 28
 [11] Bayro-Corrochano E J, Arana-Daniel N 2010 *IEEE Trans. Neural Netw.* **21** 1731
 [12] Grinblat G L, Uzal L C, Ceccatto H A, Granitto P M 2011 *IEEE*

- Trans. Neural netw.* **22** 37
- [13] Yang X W, Zhang G Q, Lu J, Ma J 2011 *IEEE Trans. on Fuzzy Syst.* **19** 105
- [14] Cristianini N, Shawe-Taylor J (Translated by Li G Z, Wang M and Zeng H J) 2005 *An introduction to support vector machines and other kernel-based learning methods* (Beijing: Publishing House of Electronics Industry) (in Chinese) [Cristianini N, Shawe-Taylor J 著李国正, 王猛, 曾华军译 2005 支持向量机导论 (北京: 电子工业出版社)]
- [15] Schölkopf B, Bartlett P, Smola A, and Williamson R *Proceedings of ICANN'98, Perspectives in Neural Computation* (Berlin: Springer L. Niklasson, M Bodén, and T Ziemke, Ed.) p111
- [16] Kwok J T, Tsang I W 2003 *IEEE Trans. Neural Netw.* **14** 544
- [17] Smola A, Murata N, Schölkopf B and Muller K 1998 *Proceedings of ICANN* (Berlin: Springer Verlag) p105
- [18] Matterna D, Haykin S 1999 *Support Vector Machines for dynamic reconstruction of a chaotic system in: Advances in Kernel Methods: Support Vector Machine* (Cambridge: MIT Press)
- [19] herkassky V, Mulier F 1998 *Learning from Data: Concepts, Theory and Methods* (New York: John Wiley & Sons)
- [20] Scholkopf B, Burges J, Smola A 1999 *Advances in Kernel Methods: Support Vector Machine* (Cambridge: MIT Press)
- [21] Vladimir C, Ma Y Q 2004 *Neural Networks* **17** 113
- [22] Cristianini N, Kandola J, Elissee A, ShaweTaylor J 2006 *J. Mach. Learn. Res.* **194** 205
- [23] Zhang S Q, Jia J, Gao M, Han X 2010 *Acta. Phys. Sin.* **59** 1576 (in Chinese) [张淑清, 贾健, 高敏, 韩叙 2010 物理学报 **59** 1576]
- [24] Rong H N, Zhang G X, Jin W D 2006 *J. Sys. Simu.* **18** 3204 (in Chinese) [荣海娜, 张葛祥, 金炜东 2006 系统仿真学报 **18** 3204]
- [25] Ljung L 1999 *System Identification-Theory for the User* (Prentice-Hall, Inc)
- [26] Zhang L F, Zhu Q M, Longden A 2009 *IEEE Trans. neural netw.* **20** 1
- [27] Mao K Z, Billings S A 2000 *Int. J. Control* **73** 132
- [28] Yu Y H, Song J D 2009 *Journal of Electronics & Information Technology* **31** 2220(in Chinese)[于艳华, 宋俊德 2009 电子与信息学报 **31** 2220]
- [29] Billings S A, Zhu Q M 1994 *Int. J. Control* **60** 1107
- [30] Prichard D, Theiler J 1995 *Physica D* **84** 476
- [31] Brock W A, Dechert W D, Scheinkman J 1996 *Econometric Review* **15** 197
- [32] Zhang L F, Zhu Q M, Longden A 2007 *Int. J. Syst. Sci.* **38** 47
- [33] Zhu Q M, Zhang L F, Longden A 2007 *Automatica* **43** 1519
- [34] Cao L J 2003 *Neurocomputing* **51** 321
- [35] Weigend A S, Huberman B A, Rumelhart D E 1990 *Int. J. Neural Systems* **1** 193
- [36] Tong H, Lim K S 1980 *J. Roy. Statist. Soc.* **42** 245
- [37] LIBSVM: a library for support vector machine Chang C C, Lin C J <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>[2011-12-5]
- [38] Ralavola L, Alche-Buc F 2003 *Proceeding of NIPS Vancouver, Canada* 2003 p981

Redundancy-test-based hyper-parameters selection approach for support vector machines to predict time series*

Yu Yan-Hua[†] Song Jun-De

(The School of Computer, Beijing University of Posts and Telecommunications, Beijing 100876, China)

(Received 18 March 2012; revised manuscript received 13 June 2012)

Abstract

The selection of hyper-parameters is a crucial point in support vector machine modeling. Different from previous method of choosing an optimal model by using basic statistics of residuals in, the new approach selects hyper-parameters by checking whether there is redundant information in residual sequence. Furthermore, omni-directional correlation function (ODCF) is used to test redundancy in residual, and the accuracy of the method is proved by theoretical analysis and numerical simulation. Experiments conducted on benchmark time series, annual sunspot number and Mackey-Glass time series, indicating that the proposed method has better performance than the recorded in the literature.

Keywords: support vector machines, hyper-parameter, residual, ODCF

PACS: 05.45.Tp, 05.45.-a, 02.50.Ey

* Project supported by the National Natural Science Foundation of China (Grant No. 61072060), the National Key Project of Scientific and Technical Supporting Programs of China (Grant No. 2009BAH39B03), the National High Technology Research and Development Program of China (Grant No. 2011AA100706), the Research Fund for the Doctoral Program of Higher Education (Grant No. 20110005120007), the Fundamental Research Funds for the Central Universities (Grant No. 2012RC0205), the Co-construction Program with Beijing Municipal Commission of Education; Engineering Research Center of Information Networks, Ministry of Education.

[†] E-mail: yuyanhua@bupt.edu.cn