基于分类 SVM 的时间序列预测研究

●毛雪岷 杨 杰

摘要:文章讨论了基于分类的 SVM 非线性回归算法及其在时间序列预测中的应用。与传统 SVM 回归算法相比,本算法有更强的不敏感性和健壮性、参数值可设定性并可避免过拟合现象。文中提出了一种计算预测模型初始参数值的方法,可以高效地找到较好的模型参数,并通过实验对方法的有效性和可行性进行了验证。

关键词:SVR(支持向量回归);时间序列;回归算法;训练算法;核函数

一、引言

预测是作为决策、规划之前的必不可少的重要环节,是科学决策、规划的重要前提。时间序列预测是预测领域内的一个重要研究方向,在过去的半个多世纪里得到了迅速的发展,特别是对线性时间序列分析的研究,已取得了系统和丰富的成果。但是,对于非线性时间序列分析的研究,仅在近二十年里才逐渐被重视起来。综观国内外在这一方向上的研究概况,前期工作大多局限于对几类典型非线性时间序列模型的参数辨识算法和建模方法等进行研究,然而,由于现实系统的复杂性,人们在预测时存在着正确选择模型的困难,便利这些方法的应用受到很大的限制。于是,人们把目光转向了近年来兴起的人工神经网络模型。传统的时间序列预测采用的是统计和神经网络等方法,如 YiMin Xiong, Di-Yan Yeung 的文章 Time series clustering with ARMA mixtures 中用的 ARMA 方法和 Ildar Batyrshin, Raul Herrera-Avelar 等的文章 Association Network in Time Series Data Mining 中使用的一种关联网络方法。统计建模方法要求时间序列具有平稳性、正态性、独立性,这个方法不适用于复杂时间序列。SVM 具有很好的非线性逼近能力,但它存在模型结构难以确定,易出现过度训练或训练不足,陷入局部最小且对连接权初值敏感,并过度依赖设计技巧。目前国外已有将支持向量机用于时间序列预测的研究,如 Sayan Mukherjee, Edgar Osuna 和 Federico Girosi 的 Nonlinear Prediction of Chaotic Time Series Using Support Vector Machines 就是这一方法的研究。但这些基于 SVM 的时间序列研究多是针对理想数据,如人工混沌序列数据等,因此支持向量机在回归中的研究还有许多不尽如人意的地方,有很大的研究余地,本文对此作了较为系统深入的研究。另外,对于现实世界中常表现出非线性时间序列,人们试图用支持向量机进行预测,但相关理论成果零星分散,且存在许多不足,本文对此进行了较深入的研究。

本文安排如下:第二节介绍了基于 SVM 的时间序列预测模型;第三节阐述了基于分类 SVM 的网络训练算法和回归算法,并将其与传统的预测模型结合;在第四节描述了一个使用这一预测方法的实验,实验数据来源:personal income and its disposition of USA: billions of dollars; SAAR(quarterly);最后在第五节分析了实验结果并对其进行了总结。

二、基于 SVM 的回归与预测

SVM 进行回归与预测的一般思想是用一个非线性映射 将数据 映射到一个高维特征空间 F 上,并在此空间进行线性回归,通过此种方法,实现将低维特征空间的非线性问题转化为高维特征空间线性回归问题解决。由统计学习理论可以得到回归函数如下:

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

此处 $\phi: R^n \to F, w \in F$。$(,)$ 表示内积,ϕ 表示 R^n 空间到 F 空间的非线性映射,$x \in R^n$,w 为权向量,$w \in F$,b 为偏置。

传统的回归问题解决方法是找到函数 f,使经验风险最小化。SVM 回归方法的思想是使得经验风险与置信风险(模型的复杂度)之和最小,使预测模型具有很好的函数逼近能力和泛化能力。式(1)中 $\phi(x)$ 已知,利用样本数据 (x_i, Y_i) 通过如下泛函最小化,可求出式(1)中的 w 和 b 估计值。

$$R_{reg}[f]=R_{emp}[f]+\lambda\|w\|^2=\sum_{i=1}^{S} C(e_i)+\lambda\|w\|^2 \tag{2}$$

这里 $R_{reg}[f]$ 为经验风险,$\|w\|^2$ 为置信风险。$C(e_i)$ 为模型的经验损失,$C(\cdot)$ 为损失函数,$e_i=f(x_i)-Y_i=Y'_i-Y_i$ 为样本预测值与真实值之差,S 为样本容量。由于 ϕ 固定,故 $\|w\|^2$ 反映了模型在高维特征空间的复杂性,该值越小则置信风险越小,λ 为用于控制样本训练损失与模型复杂性折中的正则化参数。

对于给定的损失函数,Vapnik 提出 ε 不敏感损失函数,定义为:

$$|y-f(x)|_\varepsilon=\begin{cases}|y-f(x)|-\varepsilon, & |y-f(x)|\geq\varepsilon\\ 0, & |y-f(x)|\leq\varepsilon\end{cases} \tag{3}$$

ε 用于控制回归逼近误差的宽度,控制支持向量的个数与泛化能力,其值越小,精度越高,支持向量数也越多,但泛化能力降低。采用该损失函数经验风险为:

$$R^{\varepsilon}_{emp}[f]=\frac{1}{S}\sum_{i=1}^{S}|y-f(x)|_\varepsilon$$

(2)式等价于如下优化问题:

$$\min L=\frac{1}{2}w^{T}w+C\sum_{i=1}^{S}(\zeta^{*}_{i}+\zeta_{i}) \quad (4)$$

$$\text{s.t.}\begin{cases} y_i-(w,\phi(x_i))-b\le\varepsilon+\zeta^{*}_{i} \\ (w,\phi(x_i))+b-Y_i\le\varepsilon+\zeta_i \\ \zeta^{*}_{i},\zeta_i\ge 0 \end{cases} \quad (5)$$

此处 C=1/λ，为便于求解，将原问题转换为对偶问题。

$$\max M=-\frac{1}{2}\sum_{i=1}^{S}(\alpha_i-\alpha^{*}_{i})(\alpha^{*}_{j}-\alpha_j)(\phi(x_i),\phi(x_j))+\sum_{i=1}^{S}\alpha^{*}_{i}(y_i-\varepsilon)-\sum_{i=1}^{S}\alpha^{*}_{i}(y_i+\varepsilon) \quad (6)$$

$$\text{s.t.}\begin{cases} \sum_{i=1}^{S}\alpha_i=\sum_{i=1}^{S}\alpha^{*}_{i} \\ 0\le\alpha_i\le C \\ 0\le\alpha^{*}_{i}\le C \end{cases} \quad (7)$$

解得：$w=\sum_{i=1}^{S}(\alpha_i-\alpha^{*}_{i})\phi(x_i)$，b 可由任一支持向量代入求出，得：

$$f(x)=\sum_{i=1}^{S}(\alpha_i-\alpha^{*}_{i})(\phi(x_i),\phi(x_j))+b \quad (8)$$

高维特征空间上的内积运算可定义为支持向量机的核函数：$K(x_i,x_j)=(\phi(x_i),\phi(x_j))$，只需对变量在原低维空间进行核函数运算即可得到其在高维空间上的内积，解凸二次规划问题得非线性映射为：

$$f(x)=\sum_{i=1}^{l}(\alpha_i-\alpha^{*}_{i})K(x_i,x_j)+b \quad (9)$$

由 Hilbert- Schmidt 定理知，任何满足 Mercer 条件的运算均可作为高维空间的内积。下面是最常见的几类核函数：

(1) 多项式核函数：

$K(x,x_i)=[(x\cdot x_j)+1]^{q}$ x_i, t>0, q∈S

(2) Gauss 核函数：

$K(x,x_i)=\exp\{-\frac{\|x-x_i\|}{2\sigma^{2}}\}$, σ>0

(3) Sigmoid 核函数：

$K(x,x_i)=\tanh(v(x,x_i)+c)$，v, c 是常数。

三、基于分类 SVM 的回归算法(CSVR)

1. 概述。传统的回归算法，基本上都是直接从样本数据求出回归函数，这些方法只可用于回归模型已知的情况。本节从另一个角度讨论支持向量回归问题，在新的回归算法中使用的是支持向量分类技术(SVC)而非支持向量回归技术(SVR)，这种方法文中记为 CSVR(NonLinear Support Vector Regression based on Classification)。较传统支持向量回归算法，它最大的优点是可用于非线性模型未知的情况下，这是传统的支持向量回归算法所鞭长莫及的。

2. CSVR 网络训练算法。

(1) 问题。

输入：给出一组输入样本

$[x_{i1}, x_{i2}, L, x_{in}]$, i=1, 2, L, S。

输出：期望输出是一个支持向量网络

(2) 步骤。

第 1 步将数据分为两类 $C_1=\{Z_+, X_1\}$和 $C_2=\{Z_+, X_2\}$，其中 $X_1=(x_1, y_1+\varepsilon, i=1, 2, L, S)$，$X_2=(x_1, y_1+\varepsilon, i=1, 2, L, S)$。

第 2 步通常 C_1 和 C_2 线性不可分。为此，通过一个特定的函数 φ 将两类数据映射到高维空间，并在高维空间构造两类数据 $C_1=\{Z_+, \phi(X_1)\}$和 $C_2=\{Z_-, \phi(X_2)\}$。

第 3 步 A：在特征空间上构造最优分类超平面 H，构造最小化泛函

$$\Psi(w)=\frac{1}{2}|w\cdot w|^{2} \quad (10)$$

$$\text{s.t. } z_i[(\phi(x_i)\cdot w)+b]\ge 1,\ i=1, 2, L, S \quad (11)$$

其中 w 是最优分类超平面权值

第 3 步 B：如果考虑到噪声的影响，在特征空间可能会有线性不可分的情况发生，可构造软间隔分类超平面，ζ_i 是松弛变量，正常数 C 控制对错分样本的惩罚程度。

$$\Psi(w,\zeta)=\frac{1}{2}|w\cdot w|^{2}+C(\sum_{i=1}^{l}\zeta_i) \quad (12)$$

$$\text{s.t. } z_i[(\phi(x_i)\cdot w)+b]\ge 1-\zeta_i,\ i=1, 2, L, S \quad (13)$$

$$\zeta_i\ge 0,\ i=1, 2, L, S$$

第 4 步：该优化问题的解由列拉格朗日泛函的鞍点求出，其中 α_i 为拉格朗日乘子。

$$L(w,b,\alpha)=\frac{1}{2}|w\cdot w|-\sum_{i=1}^{l}\alpha_i z_i\{[\phi(x_i)\cdot w+b]-1\} \quad (14)$$

函数 L 的极值应该满足条件

$$\frac{\partial}{\partial w}L=0,\ \frac{\partial}{\partial a}L=0,\ \frac{\partial}{\partial b}L=0,$$

最后得到原问题的对偶形式：

$$W(\alpha)=\sum_{i=1}^{S}\alpha_i-\frac{1}{2}\sum_{i=1}^{S}\sum_{j=1}^{S}\alpha_i\alpha_j z_i z_j K(x_i,x_j) \quad (15)$$

其中 $K(x_i,x_j)$ 是满足 Mercer 条件的核函数。

第 5 步：求解此对偶问题，可求出一系列拉格朗日乘子系数 α_i，一般情况下 α_i 中有很多项的值为 0，α_i 中不为 0 的项称作支持向量。

至此，我们已经构造出 CSVR 支持向量回归网络，其中使用了标准的支持向量二分类技术。

3. CSVR 回归算法。

CSVR 回归算法可简单描述如下：

(1) 问题。

输入：CSVR 支持向量回归网络和测试向量$[x_{i1}, x_{i2}, L, x_{in}]$, i=1, 2, L S。

输出：y_i, i=1, 2, L, N

(2) 步骤。

第 1 步：设向量 X =[X, Y]取自函数 Y=f(X)，将该向量输入到 CSVR 支持向量网络，则其输出一定落在最优分类超平面上，即

$$\sum_{i=1}^{N}\sum_{j=1}^{S}\alpha_i z_j K(x_i', x_j) \quad (16)$$

在式(16)中，除了包含在 x_i 中的 y_i 是未知变量，其他都是已知的，若求出 y_i，也就求出 $x_i\rightarrow y_i$，就能求出 $y_i=f(x_i)$，i=1, 2, L, N, 此即为所求回归问题。

第 2 步：由于输入向量 X 的输出值的每一点都必然落在最优分类超平面上，则式(16)可拆分成 S 个求解 y_i 的一

元方程:

$$\sum_{j=1}^{S}\alpha_j z_j K(x_i', x_j)=0,\ i=1,2,L,N \qquad (17)$$

第 3 步:由于支持向量机的特点,拉格朗日乘子系数 α_i 中包含了大量的 0, 一般情况下不为 0 的个数 l<<N,则上式可以进一步简化为 $\sum_{j=1}^{l}\alpha_j z_j K(x_i', x_j)=0,\ i=1,2,L,N$。

第 4 步:求解 y_i 有很多数值解法,如最速下降法、马尔夸特法等方法,此处采用最速下降法。令目标函数为 $F(y_i)=\sum_{j=1}^{l}\alpha_i z_i K(x_i', x_i)$,迭代计算 $y_i(j+1)=y_i-\eta\frac{dF(y_i)}{dy_i}$,求出 y_i,其中 η 为步长。

4. 算法中 ε 值和 σ 初始值的计算方法。

(1)不敏感损失函数 ε 的引入,把 SVM 推广到非线性系统的回归估计,并展现了极好的学习性能。基于 SVM 方法的回归估计以可控制的精度逼近任意非线性函数,同时具有全局最优、良好的泛化能力等优越性能。SVM 通过参数控制回归估计的精度,但 ε 取多少才能达到所期望的估计精度是不明确的。

在本文中,我们通过实验得出了一个比较有效地确定 ε 初始值的简单方法,可以通过 $(d''d')/sprt(d)=2\varepsilon$ 来计算, 其中 d 是原始样本数据的平均值,d'是原始数据各点的二阶导数的平均值。这种方法给定的 ε 初始值不是最优,只是一个大概的值,但它离最优值很接近。我们通过经验调节得到的最优值往往都在这一初始值周围。这样我们在预测时可以在这一初始值的附近调整即可得到比较满意的 ε 值。

(2)预测曲线性能的优劣主要取决于核函数,当选定核函数后,σ 值的选择对于曲线的预测效果也有着至关重要的影响。本例中而言,选择的核函数是 Gauss 核函数,若使用的核宽度相对来说比较大,会得到较为平缓的预测函数,这样得到的拟合效果不好,对应的泛化能力必然很差。相反,若使用的核宽度过小,会得到较为尖锐的高斯函数曲线,使得每个样本点都成为高斯函数的峰值,对应的泛化能力也很差。因此,应根据不同问题,针对性地选择合适的核函数形式及其参数。

尽管如此,我们通过多次实验,仍然找到了一种确定 σ 初始值的方法,可以通过 $lg\sigma/lgd\approx 5$ 来计算,其中 d 是原始数据各点的一阶导数的平均值。这种方法给定的 σ 值的不是最优,只是一个大概的值,但它离最优值很接近。在多次实验中我们通过经验调节得到的最优值总是在这一初始值的周围。这样我们在预测时可以在这一初始值的附近调整即可得到比较满意的 σ 值。

四、实验

实验使用美国个人收入时间序列数据 (Personal income: Personal Income and Its Disposition: Billions of dollars; SAAR(quarterly))。从 1947 年~2005 年,每季度一个数据。将 1947 年一季度~2000 年四季度的数据作为训练样本,2001 年一季度~2005 年四季度的 20 个数据作为检验样本。原始数据点图中“”所示(为表示方便,我们把从 1947 年~2005 年各季度值用从 0 到 23.5 表示。每点间隔为 0.1 表示一个季度。):

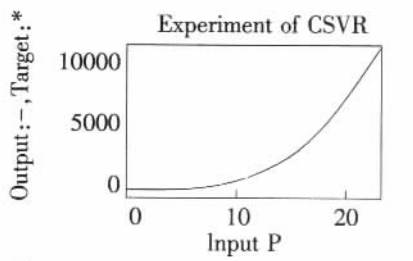

图 1 $\sigma=6.0*10^9, \varepsilon=102.15$ 的结果

实验使用 Gauss 核函数, 选择错分样本惩罚参数 C=einsensitive, $\sigma=6.0*10^9$, $\varepsilon=102.15$,实验结果如图 1 所示,图中“*”表示预测数据。从中不难看出其预测效果,不论是趋势上还是数值上都与真实值接近。

这里要指出的是:(1) 错分样本惩罚参数 C 值的确定一般是人为给定的,很难知道所取 C 值的好坏性。如何选择一个最佳的 C 值,目前在理论上尚未解决。(2)预测曲线性能的优劣主要取决于核函数。目前常用的几种核函数及其函数参数的选择都是人为的,根据经验来选取的,带有一定的随意性,因此具有局限性。在不同的问题领域,核函数应当具有不同的形式和参数,应引入领域知识,从数据依赖的角度选择核函数,这还需要进一步的研究。

五、结论

在应用于回归问题时, 支持 SVR 综合考虑了曲线平滑与误差程度,从而提高了泛化能力。本文使用了基于分类的 SVR 算法对时间序列数据进行预测。与普通的 SVR 方法相比, 基于分类的 SVR 算法可以在模型事先未知的条件下对其进行回归,并且由于是 SVR 方法,执行效率从理论上可以保证。进一步的研究包括:多维时间序列数据的预测、并尝试使用计量经济学模型解决多维时间序列的回归问题。

参考文献:

1. Yi Min Xiong, Di—Yan Yeung. Time series clustering with ARMA mixtures. Pattern Recognition, 2004,(37) :1675-1689.

2. Ildar Batyrshin, Raul Herrera-Avelar, et al. , Association Network in Time Series Data Mining, NAFIPS 2005-2005 Annual Meeting of the North American Fuzzy Information Processing Society.

3. 高隽. 人工神经网络原理及仿真实例. 北京:机械工业出版社, 2003.

4. Ye Ning, Liang Zuopeng, Dong Yisheng, Wang Huoli. SVM Nonlinear Regression Algorithm Computer Engineering, October 2005.

基金项目:本文受安徽省自然科学基金资助,项目号 070416251。

作者简介:毛雪岷,博士,合肥工业大学管理学院副教授;杨杰,合肥工业大学管理学院硕士生。

收稿日期:2007-07-13。