最小二乘支持向量机的扩展及其在时间序列预测中的应用

向小东
（福州大学管理学院，福州 350002）

［摘要］ 根据时间序列近期数据较远期数据包含有更多未来信息的思想，对最小二乘支持向量机预测方法进行了扩展，得到了更具一般性的最小二乘支持向量机预测模型，给出了扩展后的预测模型具体算法。两个时间序列的预测实例表明，扩展后的预测方法获得了更好的预测效果，提升了最小二乘支持向量机预测方法的价值。

［关键词］ 最小二乘支持向量机；扩展；时间序列；预测
［中图分类号］ G202 ［文献标识码］ A ［文章编号］ 1009－1742(2008)11－0089－04

1 前言

支持向量机是 Vapnik 在 20 世纪 90 年代提出的一种机器学习算法，它通过寻求结构风险最小化来实现实际风险最小化，从而在样本数较少时也能获得良好的学习、泛化效果。由于支持向量机出色的性能，该技术已成为机器学习领域的研究热点，并在许多领域得到了成功应用。但是，标准的支持向量机算法是一个凸二次规划问题，当样本数据量很大时，二次规划的求解会遇到很大困难。在这种情况下，Suykens 等人在 Vapnik 统计学习理论的基础上提出了最小二乘支持向量机模型[1,2]，最小二乘支持向量机将标准支持向量机的二次规划问题转换为线性方程组的问题，从而使得最小二乘支持向量机结构更简单，求解更容易，学习速度更快，并在许多领域的应用中取得了比较理想的效果。可是，进一步分析发现，一般最小二乘支持向量机在回归时仍存在一些不足，所以，文中对其进行了改进、扩展，从而得到更具一般性的最小二乘支持向量机回归模型，并用其对 Lorenz 混沌时间序列与上证综合指数时间序列进行了预测仿真，取得了不错的效果。

2 最小二乘支持向量机预测方法

最小二乘支持向量机是 Suykens 等人提出的一种新型支持向量机，用于预测的最小二乘支持向量机模型可表示为[1,2]：

$$\min J = \frac{1}{2}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} + \frac{\gamma}{2}\sum_{i=1}^{N} e_i^2 \tag{1}$$

$$s.t. \quad y_i = \boldsymbol{W}^{\mathrm{T}}\varphi(X_i) + b + e_i \quad i = 1,\cdots,N \tag{2}$$

其中，γ 为可调常数，N 为训练样本个数，b 为偏置，e_i（$i = 1,\cdots,N$）为误差。为了求解上述优化问题，建立 Lagrange 函数

$$L(\boldsymbol{W},b,e,\alpha) = J - \sum_{i=1}^{N}\alpha_i \{\boldsymbol{W}^{\mathrm{T}}\varphi(X_i) + b + e_i - y_i\} \tag{3}$$

根据 KKT 最优条件，消去 $\boldsymbol{W}$ 和 e_i 后，可得如下线性方程组

$$\begin{bmatrix} 0 & e_1^{\mathrm{T}} \\ e_1 & \Omega + \dfrac{I}{\gamma} \end{bmatrix}\begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{4}$$

其中，$\boldsymbol{y} = [y_1,\cdots,y_N]^{\mathrm{T}}$，$\boldsymbol{e}_1 = [1,\cdots,1]^{\mathrm{T}}$，$\alpha = [\alpha_1,\cdots,\alpha_N]^{\mathrm{T}}$，$\Omega = [\varphi(X_i),\varphi(X_j)] = K(X_i,X_j)$，

［收稿日期］ 2007－10－15；修回日期 2008－01－10
［基金项目］ 福建省教育厅科研基金资助(JA06022S)
［作者简介］ 向小东(1973－)，男，四川广安市人，博士，福州大学管理学院副教授，研究方向为系统预测与决策

$i,j = 1,\cdots,N$。

式中，K 为核函数。由以上可知，最小二乘支持向量机的优化问题可化为求解线性方程组问题。线性方程组的求解较二次规划问题的求解要简单快速得多。由式(4)求得 α 与 b 后，可得最小二乘支持向量机的函数估计式为

$$\hat{y}(X) = \sum_{i=1}^{N} \alpha_i K(X, X_i) + b \tag{5}$$

3 最小二乘支持向量机预测方法的扩展

在式(1)中，可调常数 γ 反映了二次误差在目标函数中的重要性，由于 γ 为常数，意味着近期数据的误差与远期数据误差同等重要。而事实上，对于时间序列来说，近期数据较远期数据包含有更多的未来信息，所以应给予近期数据的误差更大的权重(这样，为了使目标函数最小，近期数据的拟合误差就必须要小，近期数据的规律性得到强化)，从而原最小二乘支持向量机模型可变为

$$\min J = \frac{1}{2}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} + \frac{1}{2}\sum_{i=1}^{N}\gamma_i e_i^2 \tag{6}$$

$$s.t. \quad y_i = \boldsymbol{W}^{\mathrm{T}}\varphi(X_i) + b + e_i \quad i = 1,\cdots,N \tag{7}$$

通过建立 *Lagrange* 函数，并根据 *KKT* 条件，可得到线性方程组

$$\begin{bmatrix} 0 & e_1^{\mathrm{T}} \\ e_1 & \Omega + \Lambda \end{bmatrix}\begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{8}$$

其中 $\Lambda_{ij} = \begin{cases} 1/\gamma_i & \text{当 } i = j \\ 0 & \text{当 } i \neq j \text{ 时} \end{cases}$ $i,j = 1,\cdots,N$。式(6)与式(8)中的 γ_i 可按下式取值

$$\gamma_i = \gamma_0 \exp(\rho \cdot i/N + \beta) \quad i = 1,\cdots,N \tag{9}$$

式(9)中，γ_0，ρ，β 都为可调常数且 $\gamma_0 > 0$，$\rho > 0$。显然，γ_i 的取值体现了离预测期更近的样本含有未来信息更多，从而应有更大的权重的思想。而且容易看出，当参数 $\rho = \beta = 0$ 时，扩展的最小二乘支持向量机退化成一般的最小二乘支持向量机，因此，可以说一般最小二乘支持向量机是扩展最小二乘支持向量机的特殊情形，扩展最小二乘支持向量机是一般最小二乘支持向量机的推广，是一般最小二乘支持向量机的一般化。

由式(8)可解得 b 与 α 为

$$\begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 & e_1^{\mathrm{T}} \\ e_1 & \Omega + \Lambda \end{bmatrix}^{-1}\begin{bmatrix} 0 \\ y \end{bmatrix} \tag{10}$$

至此，可根据前述内容得到扩展的最小二乘支持向量机预测算法如下：

1)已知样本集 $T = \{(X_1, y_1), \cdots, (X_N, y_N)\}$，其中 $X_i \in R^m$，$y_i \in R$，$i = 1,\cdots,N$。

2)选择一个核函数 K，利用所有输入样本计算 $\Omega = K$。

3)选取参数 γ_0，ρ，β 的合适值，据式(9)计算 γ_i ($i = 1,\cdots,N$)，从而得到矩阵 Λ 的值。

4)由式(10)求得 b 与 α 的值。

5)对于一个新的输入样本，利用式(5)计算得到需要的预测值。

4 基于扩展的最小二乘支持向量机的时间序列预测

4.1 Lorenz 混沌时间序列预测

混沌时间序列预测是预测领域中较难的一个问题，下面将扩展的最小二乘支持向量机用于 *Lorenz* 混沌模型的时间序列预测。*Lorenz* 混沌模型为[3]

$$\begin{cases} \mathrm{d}x/\mathrm{d}t = -10x + 10y \\ \mathrm{d}y/\mathrm{d}t = 28x - y - xz \\ \mathrm{d}z/\mathrm{d}t = -\frac{8}{3}z + xy \end{cases} \tag{11}$$

采用四阶龙格—库塔方法得到时间序列(x)分量，时间步长为0.06，初始值为 $x = 0$，$y = 0.01$，$z = 0$，前2 000点作为暂态点去掉，把后1 000点作为原始数据。

将1 000个 x 分量的原始数据用时间序列的滞后项对时间序列进行拟合及预测(测试)，取 $m = 7$，得样本对形式 $x_i = f(x_{i-1}, x_{i-2}, \cdots, x_{i-7})$ ($i = 8,\cdots,1\,000$)，共993个样本对，其中前700个用于学习(即 $N = 700$)，剩余的293个用于预测(测试)。学习与预测(测试)中选取的核函数为三次多项式核 $K(X_i, X_j) = (1 + X_i \cdot X_j)^3$，式(9)中的参数为 $\gamma_0 = 1$，$\rho = 0.1$，$\beta = 2$(经多次试错后确定)。据前述扩展的最小二乘支持向量机预测算法，用 *matlab* 语言编程序计算得到结果如表1所示。为了与扩展前(即 $\rho = \beta = 0$)的结果进行比较，表1中也列出了扩展前的相应结果。

表1 Lorenz 时间序列拟合及预测结果

Table 1 The fitting and forecasting results of Lorenz time series

	拟合可信度/%	拟合平均相对误差/%	拟合平均绝对误差	预测可信度/%	预测平均相对误差/%	预测平均绝对误差
扩展前	77.14	1.77	0.023 5	91.81	1.07	0.014 3
扩展后	79.86	1.51	0.019 5	96.25	0.51	0.006 0

表1中,可信度指的是相对误差小于1%的样本比例,它可在某种程度上反映预测的质量,而相对误差 = |(实际值 - 估计值)/实计值|,绝对误差 = |实际值 - 估计值|。容易看出,普通的最小二乘支持向量机与扩展后的最小二乘支持向量机在拟合及预测方面都取得了比较理想的效果,它们的平均相对误差都小于2%,平均绝对误差都小于0.024。而且扩展后的最小二乘支持向量机在各方面又都要好于扩展前的情况,尤其是预测可信度、预测平均相对误差、预测平均绝对误差的改善很明显,说明采用扩展后的最小二乘支持向量机预测混沌时间序列可取得更好的效果。

4.2 股市数据预测

近几年来,股票投资受到了我国民众的普遍关注与参与,股票市场在我国得到了迅速发展。为了更好地理解股票市场以及为了获得更多的收益,股市的预测成了众多投资者及学术研究人员的热点问题。人们尝试采用了K线图模型、计量经济模型、混沌理论模型、人工神经网络模型、支持向量机模型[4~7]等进行股票市场的预测。在这众多的股市预测模型中,支持向量机模型被认为是相对最成功的。但是,由于股票市场是一个高度复杂的、高噪声的、非线性的、甚至混沌的动力学系统,受到大量相互关联的因素的影响,使得股市的预测效果还是不如人意,至今股市的预测仍被学术界看成最具挑战性的预测内容之一。另外,在流行的支持向量机股市预测方法中[4~7],人们都是采用标准的支持向量机模型,即在求解一个二次规化问题的基础上进行预测。在这里,笔者用前述扩展的最小二乘支持向量机预测方法进行股市数据的预测。

把上证综合指数看作一个时间序列,采用2006年年初到2007年8月底共402个数据进行拟合及预测(测试)。取 $m = 10$,得样本对形式 $x_i = f(x_{i-1}, x_{i-2}, \cdots, x_{i-10})$ ($i = 11, \cdots, 402$),共392个样本对,其中前300个用于学习(即 $N = 300$),剩余的92个用于预测(测试)。学习与预测(测试)中选取的核函数为一次多项式核 $K(X_i, X_j) = (1 + X_i \cdot X_j)$(考虑到时间序列具有直线上升趋势),式(9)中的参数为 $\gamma_0 = 1$, $\rho = 0.2$, $\beta = -12$(经多次试错后确定)。据前述扩展的最小二乘支持向量机预测算法,用matlab语言编程计算得到结果如表2所示。为了与扩展前(即 $\rho = \beta = 0$)的结果进行比较,表2中也列出了扩展前的相应结果。

表2 上证综指时间序列拟合及预测结果

Table 2 The fitting and forecasting results of Shanghai synthesizing index time series

	拟合可信度/%	拟合平均相对误差/%	拟合平均绝对误差	预测可信度/%	预测平均相对误差/%	预测平均绝对误差
扩展前	47	1.33	26.798 3	44.57	1.84	75.024 5
扩展后	52.33	1.29	25.438 5	48.91	1.76	71.912 1

表2中,可信度、相对误差、绝对误差的含义同表1。由表2可知,扩展前后股市数据拟合及预测的平均相对误差都在2%以内,即平均精度达98%以上,且扩展后的效果要更好一些,尤其是扩展后的最小二乘支持向量机的拟合及预测可信度的改善都比较明显,改善了4%~5%,预测的平均绝对误差也改善了约3%。所以,扩展后的最小二乘支持向量机具有更强的抗噪声能力,用其进行股市数据的预测可取得更好的预测效果。

5 结语

由于最小二乘支持向量机结构简单,学习速度快,拟合及泛化性能好,在时间序列的预测中得到了大量应用。但最小二乘支持向量机仍存在把不同时期的误差同等看待的不足,文中对这一不足进行了改进,也可理解成是对原最小二乘支持向量机的扩展,从而得到了更具一般性的最小二乘支持向量机预测模型。Lorenz混沌时间序列与上证综合指数时间序列预测仿真表明,扩展后的最小二乘支持向量机预测模型的性能较扩展前有比较明显的改善,说明扩展后的最小二乘支持向量机预测模型提升了原最小二乘支持向量机预测模型的性能,且扩展后的模型更加灵活,具有更大的价值。

参考文献

[1] 陈 磊,张土乔.基于最小二乘支持向量机的时用水量预测模型[J].哈尔滨工业大学学报,2006,38(9):1528-1530

[2] 韩 敏.混沌时间序列预测理论与方法[M].北京:中国水利水电出版社,2007,208-209

[3] 吕金虎,陆君安,陈士华.混沌时间序列分析及其应用[M].武汉:武汉大学出版社,2002.14-18

[4] 施燕杰.基于支持向量机(SVM)的股市预测方法[J].统计与决策,2005,(4):123-125

[5] 田 翔,邓飞其.精确在线支持向量回归在股指预测中的应用[J].计算机工程,2005,31(22):18-20

[6] 周万隆,姚 艳.支持向量机在股票价格短期预测中的应用[J].商业研究,2006,(6):160-162

[7] 王彦峰,高 风.基于支持向量机的股市预测[J].计算机仿真,2006,23,(11):256-258

Generalization and application in time series forecasting of the least square support vector machine method

Xiang Xiaodong

(*School of Management, Fuzhou University, Fuzhou 350002, China*)

[**Abstract**] According to the theory that the present data contains more future information than historical data in time – series, the paper extends the prediction method of least square support vector machine and obtains a more general prediction model of least square support vector machine, and develops algorithm of the extended prediction model. Prediction examples of two time – series show that the extended model is more effective. Therefore it improves the value of the prediction method of least square support vector machine.

[**Key words**] least square support vector machine; generalization; time series; forecasting