文章编号:1003-6199(2006)03-0088-04

基于时间序列的支持向量机在股票预测中的应用

彭丽芳[1],孟志青[2],姜 华[3],田 密[3]

(1.湖南工业大学图书馆,湖南 株洲 412000;2.浙江工业大学 经贸管理学院,浙江 杭州 310032;
3.湘潭大学 信息工程学院,湖南 湘潭 411105)

摘 要:由于股票预测是不确定、非线性、非平稳的时间序列问题,传统的方法往往难以取得满意的预测效果。本文提出一种基于时间序列的支持向量机(SVM)股票预测方法。利用沙河股份的股票数据,建立股票收盘价回归预测模型,该模型克服了传统时间序列预测模型仅局限于线性系统的情况。实验结果表明,该方法比神经网络方法以及时间序列方法的预测精度更高,可以很好的应用某些非线性时间序列的预测中。

关键词:支持向量机(SVM);时间序列;股票预测

中图分类号:TP181 **文献标识码**:A

Application of Support Vector Machine Based on Time Sequence in Stock Forecasting

PENG Li-fang[1], MENG Zhi-qing[2], JIANG Hua[3], TIAN Mi[3]

(1. Library, Hunan University of Technology, Zhuzhou 412000, China;
2. College of Business and Administration, Zhejiang University of Technology, Hangzhou 310032, China;
3. Department of Computer Science and Engineer, Xiangtan University, Xiangtan 411105, China)

Abstract: Because stock forecasting is a uncertain, nonlinear and nonstationary time series problem, it is difficult to achieve a satisfying prediction effect by traditional methods. This paper presents a novel stock forecasting method in which an improved Support Vector Machine (SVM) algorithm based on time sequence. Using Shahe' s stock data, a prediction model of the closing price regression is established. The model abstains from the default of traditional time series prediction model that only can be used in linear system. The experiment results are also compared with Neural networks and time sequence methods, which indicate that the SVM strategy can improve precision and therefore this prediction model can be effectively used in some nonlinear time series forecasting.

Key words: support vector machine (SVM); time series; stock forecasting

1 引 言

股票市场,具有高收益与高风险并存的特性。关于股市分析与预测的研究一直为人们所关注。但是由于股票市场高度的非线性,众多股市分析方法的应用效果都难如人意。常用的预测方法有时间序列法、灰色模型法、证券投资分析方法、专家评估法等。文献[1]采用灰色系统模型,文献[2]利用时间序列预测,但都不能很好的模拟股票预测的非线性关系。随着非线性科学的发展,人们提出了神经网络方法,通过综合系统的不确定性和工程经验,来解决复杂的设计问题。文献[3]~[5]采用经典BP神经网络进行股票预测,文献[6]~[8]对经典的BP算法进行改进,文献[6]采用的是粗神经网络,文献[7]提出了小波神经网络,文献[8]提出了进化神经网络。这些算法较经典的BP算法神经网络在收敛精度、收敛速度和全局优化方面有所

收稿日期:2005-09-09

作者简介:彭丽芳(1980—),女,湖南张家界人,硕士研究生,研究方向:数据挖掘、支持向量机(E-mail:xiaopeng427@126.com);孟志青(1962—),男,上海人,教授,研究方向:数据挖掘、运筹与管理。

改善,但这类方法存在最终解过于依赖初值,存在过学习的现象,训练过程中存在局部极小问题,且收敛速度慢,网络的隐节点难于确定等问题。

支持向量机[9-10](Support Vector Machine,简称 SVM)方法基于统计学习理论,由 Vapnik 在 90 年代中期提出。支持向量机目前已成为机器学习界的热点,成功应用于分类和回归问题。当前,SVM 已经在模式识别领域取得了很好的应用效果,广泛应用于文本识别、语音识别、人脸识别。近年来,人们发展了回归型支持向量机,它可以按任意精度逼近非线性函数,具有全局极小值点和收敛速度快的优点,被应用于天气预报[13]、地下水位预报[14]、负荷预测[15]等领域,获得了很好的效果。本文利用时间序列 SVM 方法进行股票收盘价预测,希望能为广大股票投资者提供正确、科学地把握股市动态的机会,以及及时准确的购进和抛出股票提供新的思路。

2　支持向量机的回归模型

回归分析又称函数估计,它要解决的问题是:根据给定的样本数据集$\{(x_i,y_i)\mid i=1,\cdots,n\}$,其中 x_i 为预测因子值,y_i 为预测对象值,n 为样本个数,寻求一个反映样本数据的最优函数关系 $y=f(x)$。如果所得函数关系 $y=f(x)$ 是线性函数,则称为线性回归,否则为非线性回归。SVM 的目标是寻求回归函数:

$$y=f(x)=(w\cdot x)+b \tag{1}$$

式中 w 为权重,x 为样本输入值,b 为阈值。

对于非线性问题,可以通过非线性变换将原问题映射到某个高维特征空间中的线性问题上进行求解。在高维特征空间中,线性问题中的内积运算可以用核函数来代替,即

$$K(x_i,x_j)=\phi(x_i)\cdot\phi(x_j) \tag{2}$$

核函数可以用原空间中的函数实现,没有必要知道非线性的具体形式。因此非线性问题的回归函数为:

$$f(x)=(w\cdot\phi(x))+b=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)K(x_i,x)+b \tag{3}$$

根据支持向量机回归函数的性质,只有少数 $(\alpha_i-\alpha_i^*)$ 不为零,这些参数对应的向量称为支持向量,回归函数 $f(x)$ 完全由其决定。

3　基于时间序列的数据建模

对于给定的时间序列$\{x_1,x_2,\cdots,x_n\}$,假定已知 x_{t-1} 预测 x_t,则可建立映射 $g:R^m\to R$,满足:

$$x_t=g(x_{t-1},x_{t-2},\cdots,x_{t-m}) \tag{4}$$

其中 x_t 为 t 时间的预测值,由 $x_{t-j}(j=1,2,\cdots,m)$ 而得。对时间序列$\{x_1,x_2,\cdots,x_n\}$建模,将其分成两部分,其中前 $n_{tr}(n_{tr}<n)$ 个数据用来进行训练,其余的数据用来验证模型的有效性,即为了更有效的进行预测模型的建模。

4　基于时间序列的 SVM 回归预测模型

假设 $x_i\in R^k$ 为影响预测的因素,$y_i\in R$ 为预测值$(i=1,\cdots,n_{tr}-m)$。支持向量机回归预测模型的建立就是寻找 x_i,y_i 之间的关系 $f:R^k\to R$,满足:

$$y_i=f(x_i)(i=1,\cdots,n_{tr}-m) \tag{5}$$

其中,R^k 为影响预测的因素,本文为了简便起见仅以股票收盘价作为预测因子;R 为股票收盘价预测值。

根据支持向量机理论,基于时间序列的预测模型的回归函数表示如下:

$$y_t=\sum_{i=1}^{n_{tr}-m}(\alpha_i-\alpha_i^*)K(x_i,x_t)+b \tag{6}$$

其中,x_t 为影响预测的因素,x_i 为 $n_{tr}-m$ 个样本中的第 i 个样本,$K(x_i,x_t)$ 为核函数,且 $t=m+1,\cdots,n_{tr}$。则有一步预测模型为:

$$\hat{y}_{n_{tr}+1}=\sum_{i=1}^{n_{tr}-m}(\alpha_i-\alpha_i^*)K(x_i,x_{n_{tr}-m+1})+b \tag{7}$$

其中,$x_{n_{tr}-m+1}=g(x_{n_{tr}-m+1},x_{n_{tr}-m+2},\cdots,x_{n_{tr}})$。一般的可得到 l 步预测模型为:

$$\hat{y}_{n_{tr}+l}=\sum_{i=1}^{n_{tr}-m}(\alpha_i-\alpha_i^*)K(x_i,x_{n_{tr}-m+1})+b \tag{8}$$

其中,$x_{n_{tr}-m+1}=g(x_{n_{tr}-m+1},\cdots,\hat{x}_{n_{tr}+1},\cdots,\hat{x}_{n_{tr}+l-1})$。

其中系数(α_i,α_i^*)的确定可通过求解下二次规划问题获得:

$$\min\frac{1}{2}\sum_{i=1}^{n_{tr}-m}\sum_{j=1}^{n_{tr}-m}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)K(x_i\cdot x_j)+\varepsilon\sum_{i=1}^{n_{tr}-m}(\alpha_i+\alpha_i^*)-\sum_{i=1}^{n_{tr}-m}y_i(\alpha_i-\alpha_i^*)$$

$$\text{s.t.}\sum_{i=1}^{n_{tr}-m}(\alpha_i-\alpha_i^*)=0 \tag{9}$$

$$0\leqslant\alpha_i,\alpha_i^*\leqslant C,i=1,\cdots,n_{tr}-m$$

5 应用实例

5.1 数据处理和模型参数的选择

在本实验中，我们选取沙河股份从 2002 年 3 月 14 日到 2002 年 8 月 19 日，共 104 天的数据作为样本。滑动窗口长度设为 80，且时间序列模型的变量 $m = 4$，即 $x_t = g(x_{t-1}, x_{t-2}, \cdots, x_{t-4})$，其中 x_t 为 t 时间的输入值。首先选取从 2002 年 3 月 14 日到 2002 年 7 月 22 日，共 84 天的数据作为训练数据集，即 $n_{tr} = 84$。后 20 天的数据集为测试数据集。首先利用前 84 天的数据预测后一天（即第 85 天）的股票收盘价，然后滑动窗口向后移动一天，重新训练建模预测接下来的一天的股价（余下测试数据集中的数据），依次递推直至完成。

学习样本确定后，为了降低建模误差，使用 Libsvm 库[11-12] 来做预测运算，首先需要对数据进行归一化处理，归一化方式为：$(x - x_{\min})/(x_{\max} - x_{\min})$，使每一因子的数据落入 [0,1] 区间。

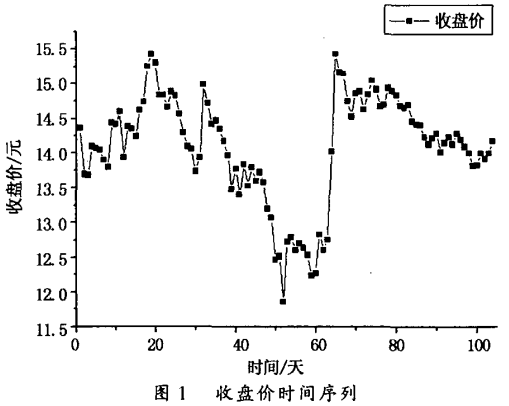

图 1 收盘价时间序列

股票预测模型的建立主要是选择相应的支持向量机参数：核函数、C 和 ε，它们对预测结果的影响很大，它们的合理确定直接影响到模型的精度和推广能力。本文通过对各种核函数的测试，最终选择径向基核函数

$$K(x_i, x_j) = \exp(-\sigma \| x_i - x_j \|^2), \sigma > 0,$$

选取 $\sigma = 0.0001$，常数 $C = 85$，$\varepsilon = 0.3$。求解 α_i，α^* 和 b，即可得到支持向量机的股票预测模型。

5.2 实验结果及分析

为了说明支持向量机的优越性，本文采用支持向量机预测模型、神经网络预测模型和时间序列预测模型分别进行提前 1 步～5 步及第 8 步和第 19 步预测。预测结果如表 1 所示。

从表 1 中可以看出，支持向量机在短期预测中具有非常理想的效果，在较长区间预测中仍然具有较高的预测精度（提前 19 步预测仍可保证平均相对误差 0.006531），虽然神经网络和时间序列在短期预测中也具有较好的效果，但在较长区间预测中推广能力降低。其中，图 2、3 分别列出了利用这三种模型提前 1 步预测和提前 19 步预测的后 20 天测试数据的预测值相对误差的比较结果。从表 1 和图 2 中可以看出，虽然基于时间序列的支持向量机模型优于另外两种模型，但这三种预测模型的预测结果的相对误差相差并不太大。从表 1 和图 3 中可以看出，支持向量机的相对误差在 0.005 范围内的有 9 个，在 0.01 范围内的有 17 个。神经网络的相对误差在 0.005 范围内的有 3 个，在 0.01 范围内的有 8 个。传统时间序列的相对误差全部都在 0.01 范围以外。因此支持向量机模型与另外两种模型相比就有了相对较大的优势，神经网络模型尤其是传统的时间序列模型与真实值之间的相对误差相差甚远，已经表现出预测能力不强。另外，表 1 所列的神经网络和时间序列预测的平均相对误差虽然只有 0.015653 和 0.033308，似乎误差不是太大，但从他们的预测曲线来看，如图 4 所示，已经明显不具有预测能力，它们表现出一种均值预测，即预测值等于训练样本的均值，而这种情况在其他的预测方法中也存在。即虽然预测值不会偏离真实值太多，计算出的误差也不会太大，但已经完全无法预测出数据的变化规律了。所以我们可以得出利用支持向量机对股票价格预测具有重要的价值。换句话说，在已知的股票价格序列基础上采用支持向量机建模，可以提前多个采样间隔时间进行有效预测，为广大股票投资者提供正确、科学地把握股市动态的机会，以及及时准确的购进和抛出股票提供指导。

表 1 预测步与平均相对误差

预测步数	平均相对误差		
	支持向量机	神经网络	时间序列
1	0.006795	0.008599	0.009419
2	0.006955	0.009500	0.010873
3	0.007129	0.010168	0.013198
4	0.006847	0.012237	0.018581
5	0.006531	0.015135	0.033308
8	0.006531	0.013882	0.033308
19	0.006531	0.015653	0.033308

注：相对误差计算公式 $\left|\frac{y_t - \hat{y}_t}{y_t}\right|$，平均相对误差计算公式 $\frac{1}{n_{te}}\sum_{t=1}^{n_{te}}\left|\frac{y_t - \hat{y}_t}{y_t}\right|$，其中 y_t 为 t 时间的实际值，$\hat{y}_t$ 为 t 时间的预测值，n_{te} 为测度数据个数。

6 结束语

股票预测受很多因素的影响，很难在股票和这

些因素之间建立一种确定的数学模型。一方面,这种关系是一种非常复杂的非线性关系;另一方面,股票价格预测还与具体的因素有很大的关系。本文分析了支持向量机用于时间序列预测的理论基础,给出了基于时间序列的支持向量机预测模型,针对沙河股份的股票数据分别采用支持向量机回归模型、神经网络模型与时间序列模型进行了预测建模和比较实验。通过实验结果分析得出:

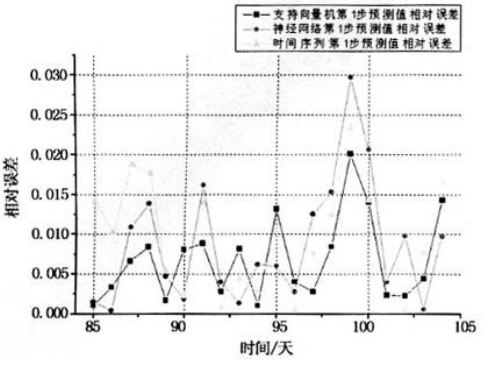

图 2　提前 1 步预测值相对误差比较图

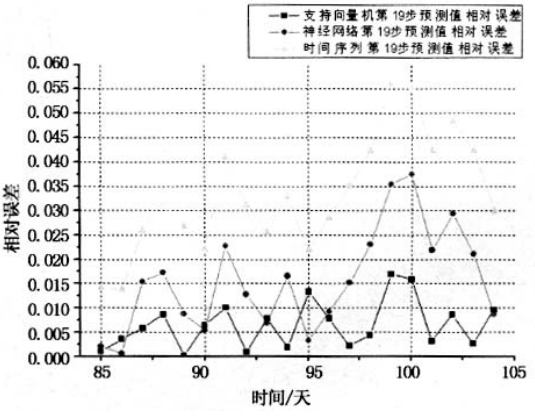

图 3　提前 19 步预测值相对误差比较图

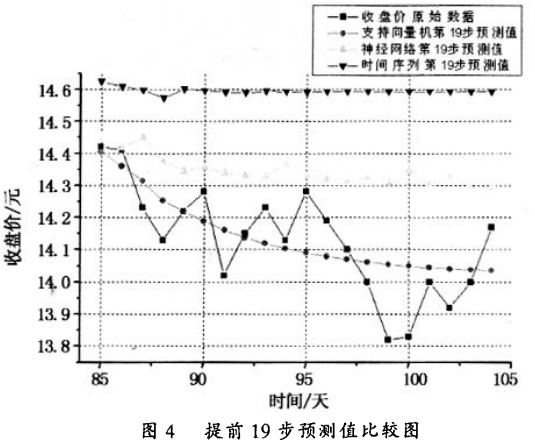

图 4　提前 19 步预测值比较图

(1)引入时间序列的支持向量机模型能够较好对股票数据进行预测。支持向量回归模型具有较快的收敛速度和计算精度,且解收敛于全局最优,使结果更接近于真实值。

(2)随着预测步数的增加支持向量机仍然具有较高的精度,这说明支持向量机具有很强的推广能力。也说明了基于结构风险最小化原理的支持向量机比基于经验风险最小化原理的神经网络有很大的优越性。

(3)支持向量回归仅取决于支持向量,而支持向量一般小于样本的个数,因而可以降低建模的复杂性。在时间序列预测中,支持向量与反映趋势变化的点紧密相关,因而回归型支持向量机能更好的跟踪时间序列的发展趋势。

(4)核函数的选择对支持向量机的学习和预测性能具有重要的影响。不同的核函数,不同的参数取值直接关系到结果的精度。

总之,采用支持向量回归的方法进行时间序列预测前景十分看好,本文分析了支持向量回归用于时间序列分析的理论基础和一般步骤,下一步则需要进一步探讨支持向量机参数选择对预测结果的影响,以便更好的为采用支持向量机进行时间序列预测提供理论指导。

参考文献

[1] 周德华. 灰色系统模型 GM(1,1)在股指灾变预测内的应用[J]. 重庆石油高等专科学校学报,2003,5(2):23-25.
[2] 谢衷洁,王驰. 用时间序列方法预测股票价格初探[J]. 数理统计与管理,2004,23(5):68-77.
[3] 吴成东,王长涛. 人工神经元 BP 网络在股市预测方面的应用[J]. 控制工程,2002,9(3):48-50.
[4] 吴贻鼎,朱翔,黄继瑜,明海山. 基于神经网络的证券市场预测[J]. 计算机应用,2002, 22(5):31-33.
[5] 蔡华. BP 神经网络与股票发行定价[J]. 计算机技术与自动化,2003, 22(1):34-39.
[6] 赵连昌,刘化波,杨艳冰,李秀艳. 一种新的粗神经网络及其在股市预测中的应用[J]. 科学技术与工程,2003,1:66-69.
[7] 吕淑萍,赵咏梅. 基于小波神经网络的时间序列预报方法及应用[J]. 哈尔滨工程大学学报,2004, 25(2):180-182.
[8] 高玮. 基于进化神经网络的股市预测研究[J]. 计算机科学,2004,31(B09):191-193.
[9] Vapnik V N.著.张学工译.统计学习理论的本质[M].北京:清华大学出版社,2000.
[10] Nello Cristianini, John Shawe-Taylor 著.李国正,王猛,曾华军译.支持向量机导论[M].北京:电子工业出版社,2004.
[11] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines[C], 2001.
[12] Chih-Wei Hsu, Chih-Chung Chang and Chih-Jen Lin, A Practical Guide to Support Vector Classification[C], 2003.
[13] 冯汉中,陈永义. 处理非线性分类和回归问题的一种新方法(Ⅱ)——支持向量机方法在天气预报中的应用[J]. 应用气象学报,2004,15(3):355-365.
[14] 王景雷,吴景社,孙景生,等. 支持向量机在地下水位预报中的应用研究[J]. 水利学报,2003,5:122-128.
[15] 潘峰,程浩忠,杨镜非,张澄,潘寇东. 基于支持向量机的电力系统短期负荷预测[J]. 电网技术,2004,28(21):39-42.