

基于 SVM 的橡胶连续收盘指数的降噪回归预测

邵明星, 王保山, 丁爱亮, 徐炳振*

(宁波大学 理学院, 浙江 宁波 315211)

摘要: 基于支持向量机的预测模型对橡胶连续的收盘价原始数据和降噪后数据进行预测, 并对拟合得到的 2 个结果进行对比. 结果表明: 对时间序列降噪后, 再用支持向量机进行预测模拟, 相比直接用原始数据得到预测结果要精确得多, 具有很好的推广价值.

关键词: 支持向量机; 降噪; 回归预测; 橡胶连续

中图分类号: O411.1; F035.4

文献标识码: A

文章编号: 1001-5132 (2012) 03-0075-04

随着动力学的发展, 人们对时间序列的复杂性有了更深刻的认识, 尤其是对时间序列的预测已经成为一个非常重要的研究方向^[1-2], 并已在信号处理、自动控制等领域中得到了广泛的应用. 近年来, 国内外都利用各种神经网络对时间序列进行预测, 并取得了成功的效果, 但是由于其采用经验风险最小化原则而大大限制了其进一步的应用. 而支持向量机是 20 世纪 90 年代中期提出的一种机器学习算法, 其理论基础是 Vapnik 提出的统计学习理论^[3]. 而统计学习理论采用结构风险最小化原则, 在最小化样本点误差的同时, 缩小模型泛化误差的上界(即最小化模型的结构风险), 从而提高了模型的泛化能力. 笔者即是基于支持向量机的理论基础, 对期货指数的收盘价进行回归分析的.

1 SVM 神经网络概述

支持向量机(Support Vector Machine, SVM)是由 Vapnik 首先提出的^[3], 像多层感知器网络和径向基函数网络一样, 可以用于模式分类和非线性回归. SVM 是基于统计学习理论发展起来的一种新的通用学习机器. 而统计学习原理采用结构风险最小化原理, 在最小化样本点误差的同时缩小模型泛化误差的上界. 此原理基于这样的事实: 学习机器测试数据的误差率(即泛化误差率), 以训练误差率和 1 个依赖于 VC (Vapnik-Chervonenkis dimen-

sion)项的和为界, 在可分模式情况下, 支持向量机对于前一项的值为零, 并且使第二项最小化^[4-5]. 而支持向量机的主要思想是通过某种事先选择的非线性映射将输入空间映射到一个高维特征空间中(图 1), 在这个特征空间中的平面上构造最优分类超平面.

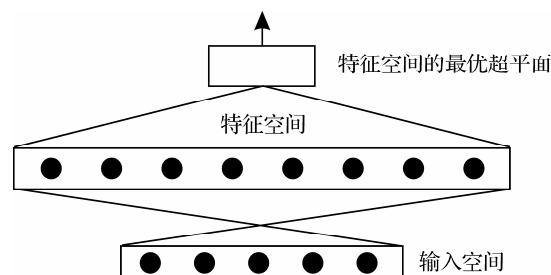


图 1 支持向量机的基本思想

用以解决函数回归预测的 SVM 表述如下: 若给定样本集 $\{x_k, y_k\}_{k=1}^N, x_k \in R^d, y_k \in R$, 所要求的拟合函数形式为:

$$y(x) = \bar{\omega} \cdot \phi(x) + b,$$

其中, $\bar{\omega}$ 为权向量; b 为偏差; $\phi(x)$ 为非线性映射函数. SVM 利用 $\phi(x)$ 将输入空间映射到 1 个高维(可能是无限维)特征空间(Hilbert 空间)中, 并在这个新空间中求取最优线性分类面, 而实现数据的线性可分; 并且 $\phi(x)$ 可通过定义, 在内积空间满足 Mercer 条件的核函数 $k(x, x')$ 来代替, 而核函数可以用原空间的函数来实现, 且无需知道 $\phi(x)$ 的

收稿日期: 2012-02-20.

宁波大学学报(理工版) 网址: <http://3xb.nbu.edu.cn>

基金项目: 国家自然科学基金(10774080).

第一作者: 邵明星(1986—), 女, 山东临沂人, 在读硕士研究生, 主要研究方向: 金融物理与复杂系统动力学. E-mail: shmx1213@126.com

*通讯作者: 徐炳振(1959—), 男, 山东德州人, 博士/教授, 主要研究方向: 金融物理与复杂系统动力学. E-mail: xubingzhen@nbu.edu.cn

具体形式.

在支持向量 x_i 和输入空间抽取的向量 X 之间的内积核个概念是构造支持向量机学习算法的关键. 支持向量机是由算法从训练数据中抽取小的子集构成的^[6]. 图 2 即为支持向量机的体系结构, 其中, K 为核函数, 其种类主要如下:

线性核函数: $K(x, x_i) = x^T x_i$;

多项式核函数: $K(x, x_i) = (\gamma x^T x_i + r)^p, \gamma > 0$;

径向基核函数: $K(x, x_i) = \exp(-\gamma \|x - x_i\|^2), \gamma > 0$;

两层感知器核函数: $K(x, x_i) = \tanh(-\gamma x^T x_i + r)$.

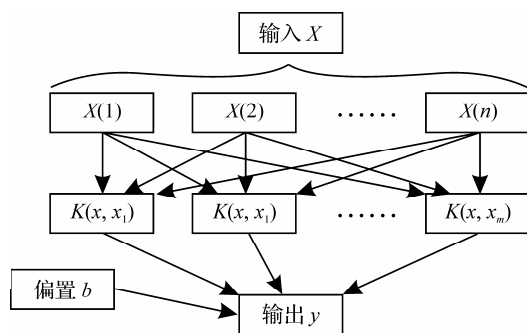


图 2 支持向量机体系结构

2 模型建立

模型建立目的是利用 SVM 建立的回归模型对期货指数的每日收盘数进行回归拟合. 以橡胶连续为例, 模型假设橡胶连续每日的收盘指数与前一日的开盘指数、指数最高值数、指数最低值数、收盘指数、交易量及交易额相关, 即将前一日的收盘指数、指数最高值数、指数最低值数、收盘指数数、交易量数、交易额作为当日收盘指数的自变量, 当日的收盘价作为因变量进行拟合.

算法流程: 根据模型假设选定因变量和自变量→数据预处理→交叉验证选择回归最佳参数→利用最佳参数训练→拟合预测→拟合预测指标.

3 实例分析

下面以橡胶连续为例进行回归预测模拟:

选定橡胶连续从 1998 年 1 月 15 日到 2011 年 5 月 31 日时期的每日开盘价、最高价、最低价、收盘价、成交量、成交额为测试数据, 一共是 3088 个交易日数据, 部分数据及相关格式见表 1.

3.1 根据模型假设选定自变量和因变量

选取第 1 个交易日到第 3087 个交易日内每日的开盘价、最高价、最低价、收盘价、成交量、成交额作为自变量, 选取第 2 个交易日到 3088 个交易日内每日的收盘价作为因变量.

3.2 数据预处理

利用 SVM 方法对橡胶连续指数进行预测模拟. 为了提高收敛速度和预测的准确率, 需要对原始的数据进行归一化处理, 将数据归一到 [1,2] 区间, 图 3 为原始橡胶连续指数和降噪后橡胶连续指数每日收盘价归一化后的结果.

但是对于原始数据来说, 一般都存在着很大的噪声, 因为噪声的存在导致预测结果的不准确, 基于此原因, 笔者将原始数据降噪, 再对降噪后的数据进行归一化. 需要说明的是这里不但需要对因变量(橡胶连续的每日收盘价)进行降噪, 还要对于自变量也需要做同样的预处理.

3.3 参数选择

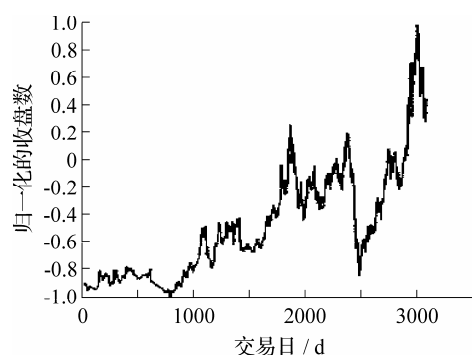
关于 SVM 参数的优化^[7], 目前常用方法是将 c 和 g 在一定范围内取值. 对于取定的 c 和 g 将训练集作为原始数据集, 利用 K-CV 方法得到在此组 c 和 g 下训练集验证的分类准确率, 最终使训练集验证分类准确率最高的那组 c 和 g 作为最佳参数.

3.4 训练和回归预测

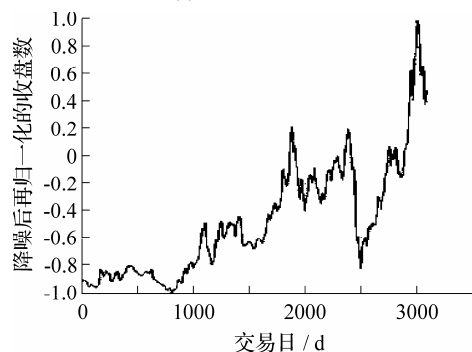
利用以上方法得到 2 组数据的最佳参数 c 和 g

表 1 橡胶连续指数的测试数据

交易日期	开盘价/元	最高价/元	最低价/元	收盘价/元	成交量/手	成交额/手
1998-01-15	7480	7540	7380	7470	5422	23996
1998-01-16	7400	7550	7400	7550	11394	23938
1998-01-19	7610	7800	7600	7710	10502	14598
1998-01-20	7680	8000	7680	7940	2840	17156
1998-01-21	7980	8100	7580	7710	2662	15060
1998-01-22	8200	8210	7800	7910	2492	17364
...	...	...	...	...	...	...



(a) 原始数据源



(b) 降噪数据源

图3 橡胶连续每日的收盘归一化结果

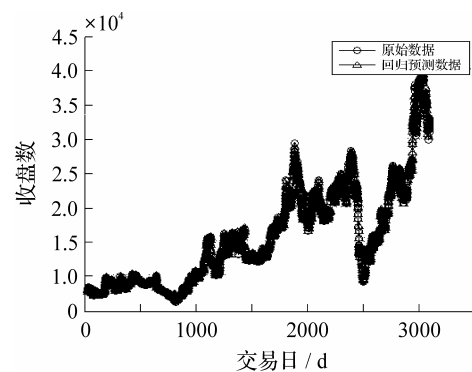


图4 原始数据回归预测图

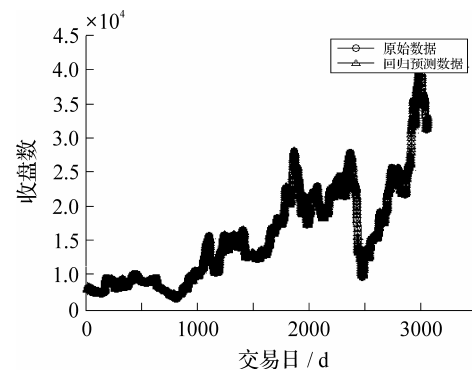


图5 降噪后数据回归预测图

对 SVM 进行训练, 然后再分别对 2 组数据进行回归预测, 最终得到的回归预测结果(图 4 和图 5)。

通过图 4 和图 5 的比较可以看出, 对原始数据回归预测和对降噪后数据回归预测得到的走势图吻合程度都非常高, 说明 2 种方法都能够较好地对期货指数进行预测, 但是具体哪一种方法能够更加接近于真实的期货指数的走势, 还需进一步进行分析。图 6 和图 7 分别为 2 种方法的相对误差比较, 通过对比可以发现, 橡胶连续的原始数据的相对误差基本都集中在区间 $[-0.05, 0.05]$ 内, 而降噪后的橡胶连续收盘指数数据的相对误差则是大多都集中在 $[-0.01, 0.01]$ 区间内, 后者比前者的误差小了整整 5 倍, 因此可知, 降噪后数据的预测结果更接近于实际价格走势, 也说明降噪后再用支持向量机进行价格预测的准确度更高^[8]。

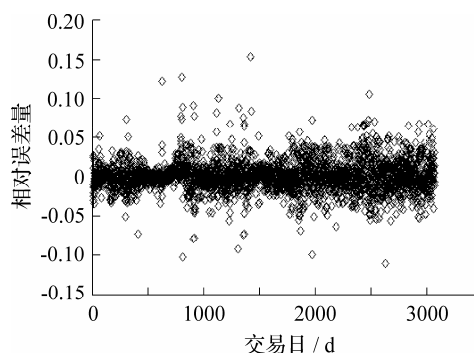


图6 原始数据结果的相对误差图

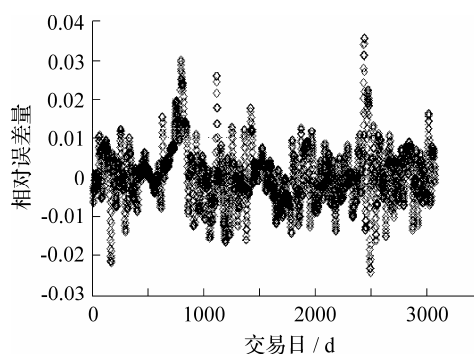


图7 降噪后数据结果的相对误差图

表 2 为 2 种方法的指标量对比, 由对比结果可以看出, 降噪后橡胶连续收盘价的预测结果的均方相对误差要小的多, 拟合曲线相关系数也更加接近于 1, 所以同样可以得出降噪后再用支持向量

表2 橡胶连续指数预测结果表

方法	均方相对误差(MSE)	相关系数(R)/%
降噪前	3.536×10^{-4}	99.780 5
降噪后	3.835×10^{-5}	99.977 8

机进行价格预测走势准确度更高的结论.

4 结语

期货指数走势一直都是人们关注的焦点,针对期货指数走势的不确定性,提出了支持向量机的期货指数预测模型^[9],并对这一模型进行了改进,并且以橡胶连续指数为例进行了预测,预测结果也表明了降噪后的支持向量机的预测结果在一定程度上更加精确,这一结果对于期货指数的预测有着重要的理论和实践意义.

参考文献:

- [1] Farmer J D, Sidorowich J. Predicting chaotic time series [J]. Phys Rev Lett, 1987, 59:845-848.
- [2] 甘建超, 肖先赐. 基于相空间邻域的混沌时间序列自适应预测滤波器(II)非线性自适应滤波[J]. 物理学报, 2003, 52(5):2995-2999.
- [3] Vapnik V N. 统计学习理论的本质[M]. 张学工, 译. 北京: 清华大学出版社, 2000.
- [4] 曹丽娟, 王小明. 金融工程的支持向量机方法[M]. 上海: 上海财经大学出版社, 2007.
- [5] Cortes C, Vapnik V. Support vector network[J]. Machine Learning, 1995, 20: 273-297.
- [6] 徐建华, 张学工, 李衍达. 支持向量机的新方法[J]. 控制与决策, 2004, 19(5):481-485.
- [7] 冯振华, 杨杰明. SVM 回归参数探讨[J]. 机械工程与自动化, 2007(3):17-22.
- [8] Cao L, Hong Y, Fang H, et al. Predicting chaotic time series using wavelet network[J]. Physica D, 1995, 85(1): 225-238.
- [9] Atsalakis G, Valavanis K. Surveying stock market forecasting techniques, Part II: Soft computing methods[J]. Expert Systems with Application. 2009, 36:5932- 5941.

On Consecutive Closing Price of Rubber Based on SVM's Regression Forecast with Noise-reduction

SHAO Ming-xing, WANG Bao-shan, DING Ai-liang, XU Bing-zhen*

(Faculty of Science, Ningbo University, Ningbo 315211, China)

Abstract: Based on support vector machine, the raw data and data with noise reduction are separately used to predict rubber consecutive closing price, and comparison of these two results is conducted and conclusion is obtained. The study suggests that the forecast using the data of time series with noise reduction is more accurate than the one directly with the raw data.

Key words: support vector machine; noise reduction; regression prediction; rubber continuous

(责任编辑 章践立)