基于最优小波包变换、ARIMA 与 SVR 的股票价格预测研究

高　天

(中央财经大学 保险学院,北京　100081)

摘　要: 股票价格序列的变化往往具有高度的非平稳性和异方差性,使得单一的预测方法难以准确预测。利用最优小波包变换,将股票价格序列分解为一系列特征规律较明显的小波包系数,对其中的趋势部分采用 ARIMA 进行预测,对细节部分采用 SVR 进行预测,最后将预测结果进行重构得到股价预测序列。实证研究结果表明: 该预测方法结构明确,计算高效,能够以较高的精度对股价变化进行预测。

关键词: 最优小波包变换; ARIMA; SVR; 股票价格; 预测

文章编号: 2095 - 5960(2015)06 - 0057 - 13; 中图分类号: F830.91; 文献标识码: A

一、引言

股票价格序列的变化由于受到整个外在经济环境、政策法规和公司经营状况和其他不可预知因素的影响,往往具有高度的非平稳性和异方差性,这使得股票价格序列这种复杂的金融时间序列难以通过一些直观的方法进行预测。在这种情况下,学术界尝试使用各种非线性的技术来对股票价格序列进行建模。目前用来预测股票价格的方法主要有 ARMA、ARIMA、GARCH 等时间序列回归方法[1][2][3]和人工神经网络、支持向量机等智能算法[4][5][6]。但是对于股票价格这种复杂的金融时间序列简单地采用一种方法来预测很难得到满意的结果,究其原因主要是股价序列波动过于剧烈,趋势性不够明显。

针对这个问题,一些文献中提出使用混合模型来进行股价预测,比如使用 ARIMA 和 RBF 神经网络[7]、ARIMA 和 SVM 相结合[8]。但这些方法只是对预测方法进行改进,并没有对数据本身进行处理,依然没有解决股价序列的高度非线性问题。有的文献提出使用小波变换对股价序列先进行分解,然后再使用各种方法进行预测[9][10][11],小波分析有数学显微镜之称,它对于复杂信号的分解非常有效,这些研究提高了对于股价预测的精度,但是在进行小波变换时仅对趋势部分进行了分解,对由此产生的细节部分则没有有效的分析。有的文献提出使用小波包基对股价序列趋势和细节部分同时进行分解[12],然而与不同的股价序列特性对应的小波包基并不一样,而且分解细节部分过多使得在分解层数高时计算量非常大且会降级预测精度。因此本文提出了一种基于最优小波包变换、ARIMA 和 SVR 的预测方法,该方法首先对股价序列进行多层分解,其次使用信息熵代价函数寻找到对应于股价序列的最优小波包基,再次对分解得到的趋势系数使用 ARIMA 进行回归预测,对细节系数进行白噪声检验,通过检验的细节系数使用 SVR 进行回归预测,最后将预测得到的各部分系数通过小波包重构合成最终的预测结果。

二、基于最优小波包变换、ARIMA 与 SVR 的股票价格预测模型

股票价格由于受到政治经济社会心理等各方面因素的影响,往往具有高度的非平稳性和异方差性,

收稿日期: 2015 - 08 - 18

作者简介: 高　天(1987—),男,贵州贵阳人,中央财经大学保险学院博士研究生,研究方向为量化投资与金融建模。

直接使用上述各种预测方法往往效果很差，特别是容易产生“平移现象”（由于股价波动太剧烈，使得预测模型对某时点的预测仅仅是单纯复制上个时点的观察值，从图形上看起来就像序列滞后一期平移一样）和“放弃预测”（同样由于股价波动的高度非平稳性和异方差性，使得某些单一方法“跟不上”股价的变动，从而使得这些方法的预测值往往在一个值附近小幅波动，看起来就好像没有预测一样，如图 1 所示），因此需要对股价时序进行特征提取。使用差分的方法是一种趋势提取的选择，但是股价的变动含有许多无法预测的白噪声，并且波动非常频繁，使得直接使用差分的方法效果并不好，因此需要对股价时间序列进行预处理，提取股价线性特征和非线性特征，分别对线性特征和非线性特征采用最合适的预测方法。

本文首先使用最优小波包变换，将股价序列分解成一个低频趋势部分和若干个高频细节部分。由于信息熵代价函数的引入，使得最优小波包变换相比于小波包变换方法分解出的各部分在信息熵意义下包含最多的有效信息，能够提升预测的精度。其中趋势部分体现了股价的整体变动趋势，由于将低频部分和高频部分进行分解，使得低频趋势部分不再具有高度的波动性，在这种情况下考虑使用 ARIMA 这种成熟的时间序列预测方法对低频趋势部分进行预测，可得到股价变动总体变动方向。高频细节部分由于受到各方面短期影响因素的影响，其中一部分由于高度的波动性和异方差性，分解后仍然无法从中提取有效的预测信息，因此考虑将其剔除。在本文中使用 LBQ 随机白噪声检验进行筛选，对没有通过 LBQ 检验而被认定为白噪声的细节系数剔除出预测范围，剩下的高频细节系数集直接使用传统的时间序列分析方法效果很差，因此考虑使用能够有效处理高度非线性和小样本容量特性问题的统计学习方法来进行高频细节部分的预测。在本文中使用支持向量回归机（Support Vector Mechine for Regression，下称 SVR）对细节系数进行预测，SVR 是有监督统计学习方法的一种，由于基于严格的 VC 维理论，能够使得预测的结构风险最小化，因此相比于传统预测方法，在对于高度非线性和异方差性问题的处理上更具优势。最后得到了各部分的预测序列，使用对应与最优小波包分解方法的最优小波包重构，将各部分预测序列重构为原始股价序列的形式，就得到了股价预测序列，组合模型整体框架如图 1 所示。

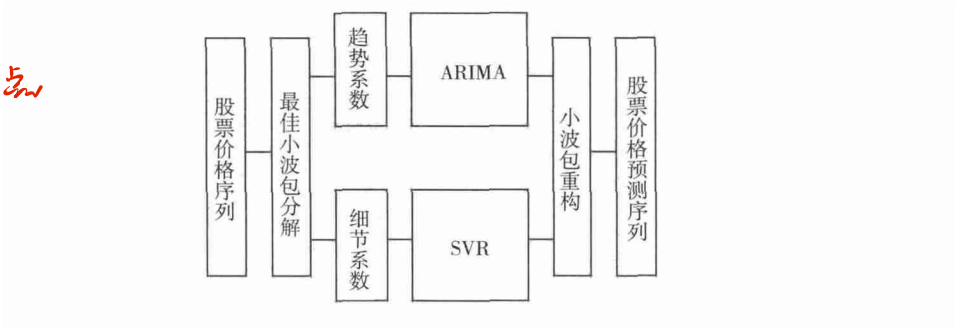

图 1 基于最优小波包变换、ARIMA 和 SVR 的股票价格预测方法步骤图

三、最优小波包变换

（一）小波变换

小波变换[13]这一概念的首次提出是由法国的石油工程师 J. Morlet1974 年在研究利用人造地震来探明原油储量时提出的一种地震回波信号解析变换方法。1986 年数学家 Y. Meyer 首次构造出了一个真正的小波基之后，小波分析才真正开始得到学术界的重视并迅速得到发展，其中比利时女数学家 I. Daubechies 撰写的《Ten Lectures on Wavelets》[14]对小波变换在学术界的普及起了重要的作用。小波变换是一个时域上和频域上的局部变换，通过伸缩变换和平移变换等运算对函数或信号进行多尺度分析，它的出现弥补了 Fourier 变换无法在时频域局部展开获得细节的缺陷，同时具有数学上严格意义的突变点诊断能力，从而小波分析技术被称之为“数学显微镜”，它是信号分析发展史上的重要里程碑。目前小波分析在信号处理、图像压缩、语音编码、模式识别、地震勘探、大气科学、金融建模以及许多非线性研究领域内得到了广泛的应用。

小波(Wavelet)函数的数学定义是:设 $\psi(t) \in L^2(R)$

若其 Fourier 变换 $\hat{\psi}(\bar{w})$ 满足: $W_\psi = \int_R \frac{|\bar{\psi}(\bar{w})|^2}{|w|} dw < \infty$ 时,则称 $\psi(t)$ 为小波母函数,并称上式是小波函数的可容许条件。

将小波母函数 $\psi(t)$ 进行伸缩和平移,设其尺度系数为 a,小波系数为 b,记变换后的函数为 $\psi_{a,b}(t)$,则:

$$\psi_{a,b}(t) = |a|^{-\frac{1}{2}} \psi\left(\frac{t-\tau}{a}\right), a, b \in R; a \neq 0$$

称 $\psi_{a,b}(t)$ 为参数 a 和 b 的小波基函数,它们是由同一母函数 $\psi(t)$ 经伸缩和平移后得到的一组函数系列。

金融市场数据大多是以离散信号形式存放的,所以需要将连续小波变换离散化才能够应用到金融时间序列分析中。需要注意的是这里所说的离散化都是针对尺度系数 a 和平移系数 b。一般来说,令:

$a = \frac{1}{2^j}, b = \frac{k}{2^j}, j, k \in Z$,则有 $\psi_{a,b}(t) = 2^{j/2}\psi(2^j t - k)$

也写作 $\psi_{j,k}(t)$。为了能重构信号 $f(t)$,要求 $\{\psi_{j,k}\}_{j,k \in Z}$ 是 $L^2(R)$ 的 Riesz 基。

(二)多尺度分析

著名数学家 Mallat 提出了多分辨分析(Multiresolution Analysis,MRA)的概念,统一了以前的各种具体小波基的构造方法,更重要的是,Mallat 多分辨率分析的框架,提出了现今广泛使用的 Mallat 快速小波变换算法,空间 $L^2(R)$ 的多分辨分析是指构造该空间内一个子空间列 $\{V_j\}_{j \in Z}$,使其具有以下性质:

(1) 单调性:

$$\cdots \subset V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset V_{-2} \subset \cdots$$

(2) 逼近性:

$$close\{\bigcup_{j=-\infty}^{\infty} V_f\} = L^2(R), \bigcap_{j=-\infty}^{\infty} V_f = \{0\}$$

(3) 伸缩性:

$$\phi(t) \in V_j \Leftrightarrow \phi(2t) \in V_{j-1}$$

(4) 平移不变性:

$$\phi(t) \in V_j \Leftrightarrow \phi(t - 2^{j-1}k) \in V_j, \forall k \in Z$$

(5) Riesz 基存在性:

存在 $\phi(t) \in V_0$,使得 $\{\phi(2^{-j}t - k)\}_{k \in Z}$ 构成 V_j 的 Riesz 基。

则称 $\phi_{j,k} = 2^{-j/2}\phi(2^{-j}t - k)$, $k \in Z$ 为尺度函数,特别地,若 $\{\phi(2^{-j}t - k)\}_{k \in Z}$ 构成 V_j 的标准正交基,则称 $\phi_{j,k}$ 为正交尺度函数。

(三)最优小波包变换

首先引入小波包基,设 $\{h_n\}_{n \in Z}$ 为正交尺度函数 $\mu_0(t)$ 对应的低通滤波器,$\{g_n\}_{n \in Z}$ 为正交小波函数 $\mu_1(t)$ 对应的高通滤波器,并有 $g_n = (-1)^n h_{1-n}$,则由:

$$\begin{cases} \mu_{2n}(t) = \sqrt{2}\sum_{k \in Z} h_k \mu_n(2t - k) \\ \mu_{2n+1}(t) = \sqrt{2}\sum_{k \in Z} g_k \mu_n(2t - k) \end{cases}$$

定义的函数 μ_n, $n = 0, 1, 2, \cdots$ 称为由正交尺度函数 $\mu_0 = \phi$ 所确定的小波包基库。

令 U_j^n 表示由 $2^{j/2}\mu_n(2^j t - k)$ 的线性组合而成的子空间,则有:

$$U_{j+1}^n = U_j^{2n} \oplus U_j^{2n+1}, j \in Z$$

则子空间 U_3^0 的三层小波包分解如图 2 所示:

小波包基库中的一组正交基称为小波包基，例如图 3 中正交的阴影部分，不同的子空间时域和频域特征对应于不同的小波包基，这些小波包分解得到的不同部分系数应当具有显著的差异并能够充分反映不同部分的特点，所以如果部分系数的差异性很大，则这部分的系数就能够充分刻画原始的时间序列，因此应该选择一种最优的小波包基，使时间序列的特性体现在尽可能少的系数上。在本文中我们通过设定信息熵代价函数来选择，某小波包基对应的熵值最小，该小波包基就是最优小波包基。

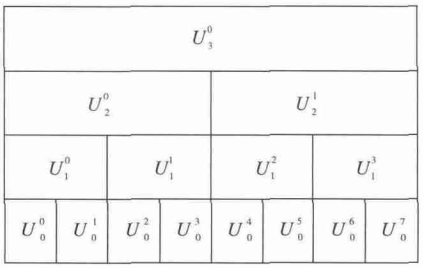

图 2 三层小波包分解示意图

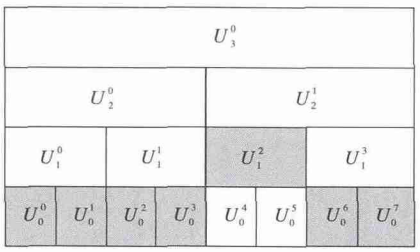

图 3 三层最优小波包分解示意图

在一个正交小波包基下将原序列展开，使得原序列对应一个小波包系数序列 $s=\{s_i\}$，则信息熵代价函数可以定义为：

$$E(s) = -\sum_i s_i^2 log(s_i^2)$$

然后求出使以上信息熵代价函数最小的正交小波包基即可得到最能反映原序列特征的最优小波包基。

四、ARIMA 时间序列回归模型

(一) ARIMA 模型数学描述

当某金融时间序列通过白噪声检验后，便可以进行回归预测，标准的处理方法是使用求和自回归移动平均模型(ARIMA)[15]对时间序列进行回归预测。

把具有如下结构的模型称之为自回归移动平均模型或 ARIMA(p,d,q)：

$$\begin{cases} \Phi(B)\nabla^d x_t = \Theta(B)\varepsilon_t \\ E(\varepsilon_t)=0, Var(\varepsilon_t)=\sigma_\varepsilon^2, E(\varepsilon_t\varepsilon_s)=0, s\neq t \\ E(x_s\varepsilon_t)=0, \forall s<t \end{cases}$$

其中：

$\nabla^d = (1-B)^d$，

$\Phi(B) = 1-\phi_1 B-\cdots-\phi_p B^p$，为 p 阶自回归的系数多项式，

$\Theta(B) = 1-\theta_1 B-\cdots-\theta_q B^q$，为 q 阶移动平均系数多项式。

上式也可以记作：

$$\nabla^d x_t = \frac{\Theta(B)}{\Phi(B)}\varepsilon_t$$

(二) ARIMA 模型时间序列回归流程

ARIMA(p,d,q) 模型的原理其实是进行 d 阶差分运算后的 ARMA(p,q) 模型，这说明当时间序列非平稳时可以通过求差分运算来使得时间序列平稳化，从而对差分后的时间序列使用 ARMA 模型进行回归预测。使用 ARIMA 模型进行时间序列建模回归分为以下几个步骤：

1. 数据预处理。对待处理时间序列数据进行预处理，主要包括缺失值的处理。

2. 检验时间序列的平稳性。若时间序列平稳可直接采用 ARMA 模型进行回归预测，若非平稳则需要进入第三步进行差分处理。通常可以采用单位根检验来判定是否平稳，例如可以使用 DF 检验和 ADF

检验。

3. 差分运算。若时间序列非平稳，则需要进行差分处理，可依次从一阶差分到高阶差分来比较，选择使得差分序列滞后自相关图和偏自相关图能够明确定阶的最小差分阶数。

4. 自回归移动平均定阶。观察自相关图和偏自相关图，若自相关图 q 阶滞后截尾，偏自相关图 p 阶滞后截尾，则可选择 ARIMA(p，d，q) 模型进行回归。

5. 进行回归，并对回归拟合结果进行显著性检验。主要包括: (1) 对回归质量的检验，比如 AIC、BIC、SIC 值; (2) 对回归系数显著性进行检验; (3) 对残差是否为白噪声进行检验，若残差判定为白噪声，则原时间序列有效趋势信息提取完毕，一般可使用 Ljung - Box Q 检验; (4) 对残差自相关图和偏自相关图进行观察，确认残差是否可认为是白噪声。

6. 模型比较。使用不同的参数进行回归，然后选择效果最好的作为最终的预测模型，一般可选择使 AIC、BIC、SIC 最小，且通过上述回归结果检验的模型。然后通过该最佳模型预测出的预测值即为 ARIMA 模型的预测结果。

五、支持向量回归机(SVR)

(一) 非线性 SVR 数学描述

支持向量回归机[16] (Support Vector Machine for Regression，SVR 或 SVM - R) 是一种结构风险最小化的统计学回归方法。支持向量回归机在有限样本的情况下能够在模型复杂度较低的前提下实现较高的推广能力和预测精度。SVR 通过解一个凸二次规划问题来获得最优决策函数，因此 SVR 最后得到的是全局最优解，有效地避免了在一些传统机器学习方法中收敛于局部最优解的固有缺陷。另外 SVR 对于高度非线性的回归问题，相比于其他方法，具有更好的处理能力。SVR 通过将输入的数据从欧式空间变换到高维的希尔伯特特征空间中，然后在高维希尔伯特特征空间中构造一个线性形式的决策函数来解决在欧式空间中难以直接解决的高度非线性回归问题。它更是通过创造性的引入核函数避免了在高维希尔伯特特征空间中高维运算的复杂性，使得回归算法的复杂度与维数无关，从而避免了“维数灾难”。

非线性支持向量回归问题的数学描述如下:

给定训练集 $T = \{(x_1, y_1), \cdots, (x_l, y_l)\} \in (R^n \times R)^l$，其中 $x_i \in R^n, y_i \in R, i = 1, \cdots, l.$ 根据这个训练集在 R^n 上寻找一个最优的决策函数 $g(x) = (w \cdot \Phi(x)) + b$，使得 $\hat{y}_s = g(x_s)$ 为任意输入数据 x_s 对应输出 y_s 的预测值，其中 $\Phi(x)$ 为从 R^n 欧氏空间到 Hilbert 特征空间的映射。此问题可通过构造一个二次凸规划问题来求解:

$$\min_{w,b} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*),$$

$$\text{s.t.} \quad \begin{cases} (w \cdot \Phi(x_i)) + b - y_i \leqslant \varepsilon + \xi_i, i = 1, \cdots, l \\ y_i - (w \cdot \Phi(x_i)) - b \leqslant \varepsilon + \xi_i^*, i = 1, \cdots, l \\ \xi_i^{(*)} \geqslant 0, i = 1, \cdots, l \end{cases}$$

其中，C 是惩罚函数，$\xi^{(*)} = (\xi_1, \xi_1^*, \cdots, \xi_l, \xi_l^*)^T$ 但是该问题无法直接求解，因此根据二次凸规划的性质，通过引入 Lagrange 函数来寻找上述原最优问题的对偶问题:

$$L(w, b, \xi^{(*)}, \eta^{(*)}, \alpha^{(*)}) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) - \sum_{i=1}^{l}(\eta_i\xi_i + \eta_i^*\xi_i^*) - \sum_{i=1}^{l}\alpha_i(\varepsilon + \xi_i + y_i - (w \cdot \Phi(x_i)) - b) - \sum_{i=1}^{l}\alpha_i^*(\varepsilon + \xi_i^* - y_i + (w \cdot \Phi(x_i)) + b)$$

其中 $\alpha^{(*)} = (\alpha_1, \alpha_1^*, \cdots, \alpha_l, \alpha_l^*)^T$，$\eta^{(*)} = (\eta_1, \eta_1^*, \cdots, \eta_l, \eta_l^*)^T$ 为 Lagrange 乘子向量。

不难证明，原问题的对偶问题为:

$$\min_{\alpha^{(*)}\in \mathbf{R}^l} \quad \frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(\Phi(x_i)\cdot\Phi(x_j)) + \varepsilon\sum_{i=1}^{l}(\alpha_i^* + \alpha_i) - \sum_{i=1}^{l} y_i(\alpha_i^* - \alpha_i)$$

$$s.t. \quad \sum_{i=1}^{l}(\alpha_i^* - \alpha_i) = 0, 0 \leqslant \alpha_i^{(*)} \leqslant C, i = 1,\cdots,l.$$

通过解这个最小化问题得到的 $\bar{\alpha}^{(*)} = (\bar{\alpha}_1, \bar{\alpha}_1^*, \cdots, \bar{\alpha}_l, \bar{\alpha}_l^*)^T$ 便是支持向量的系数，然后我们便可以构造出最优决策回归函数:

$$y = g(x) = \sum_{i=1}^{l}(\bar{\alpha}_i^* - \bar{\alpha}_i)K(x, x_i) + b$$

得到最优决策回归函数，便可以进行回归预测了。

(二) 核函数

从上述回归问题可以发现，变换 Φ 的使用只有在计算内积 $\Phi(x_i)\cdot\Phi(x_j)$ 时才会用到，在 Hilbert 维数很高时这种内积运算的时间复杂度非常高，因此有没有什么方法能够直接计算 $\Phi(x_i)\cdot\Phi(x_j)$ 呢，这便是核函数:

$$K(x, x') = \Phi(x)\cdot\Phi(x')$$

核函数的引入使得原本复杂的 Hilbert 高维特征空间的向量内积运算能够被迅速地计算出来，这使得支持向量机得以摆脱高维运算带来的巨大运算量，避免了“维数灾难”，随着 20 世纪末 21 世纪初计算机高性能运算能力的急剧提升，使得支持向量机迅速地从理论走向了实践，得以在许多领域得到应用，并取得了相当好的效果。

核函数一般来说有:

1. 多项式核函数: $K(x, x') = (x\cdot x' + g)^d$，其中 d 为多项式的阶数。

2. Gauss 径向基核函数: $K(x, x') = \exp(-\|x - x'\|^2/\sigma^2)$，其中参数 σ 为 Gauss 径向基核函数的待定参数，一般来说，当无法确定核函数类型时，选择 Gauss 径向基核函数作为 SVR 的核函数来进行回归，结果都不会太差。

另外还有指数径向基核函数、神经网络多隐层感知核函数、B 样条插值核函数等，不同的核函数对于回归结果有着显著地影响，甚至直接影响到回归结果的好坏和推广泛化的能力。因此在回归的过程中，一个最大的难点就是选择合适的核函数。但遗憾的是，至今为止并没有行之有效的核函数选择方法，只能够通过分别使用不同的核函数来进行回归，并选择其中表现最好的核函数作为最终的核函数形式。

此外，在使用 SVR 进行时间序列回归时，模型中有许多待定参数，比如惩罚系数 C、ε 参数、多项式核函数参数中的阶数 d 和 Gauss 径向基核函数中的参数 σ 等，不同的参数值对于最后的回归结果好坏有着显著地影响，因此这也是运用支持向量机中的另外一个难点，即模型参数的寻优，目前常用的支持向量机参数寻优算法主要有: 遗传算法(Genetic Algorithm，GA)、粒子群优化算法(Particle Swarm Optimization，PSO)、模拟退火算法(Simulated Annealing，SA) 和 SMO 优化算法(Sequential Minimal Optimization，SMO) 等。

六、基于最优小波包变换、ARIMA 与 SVR 的股票价格预测步骤

以下以 4 层小波包分解系数为例，描述组合预测的步骤:

1. 首先对原始股价序列进行预处理，剔除其中的缺失值，并且考虑到股价的特性，对股价序列进行对数变换。然后对股价序列进行最优小波包分解，选择的小波包基标准是使信息熵代价函数达到最小，在 4 层小波包分解下能够得到的系数编号按照二叉树的结构(如图 3) 从左到右，从上到下顺序编号。趋势系数的编号是 c_{15}，而其他的细节系数由于股价序列的选择不一样可能产生的细节系数编号也不一样。

2. 对分解出的趋势系数 c_{15} 来说其趋势性明显，且在经过四层分解后，其序列剔除了高频细节部分的干扰，趋势性相较于未分解前的原股价时间序列更为明确，采用 ARIMA(p，d，q) 模型对 c_{15} 进行拟合预测。

首先观察自相关图和偏自相关图采取逐阶实验的方法来选择 c_{15} 的差分阶数。然后观察自相关图和偏自相关图对差分后的趋势系数进行自回归和移动平均定阶，然后进行拟合预测，对拟合的残差序列进行 LBQ 随机白噪声检验，若残差序列通过 LBQ 检验，说明原始序列的信息基本提取完毕。

3. 对于分解出的高频细节系数集 $\{c_k\}$，$k \in S$，其中 S 为经过最优小波包变换得到的除了趋势系数 c_{15} 之外的细节系数集，首先使用 LBQ 检验对 $\{c_k\}$，$k \in S$ 进行随机性检验，未通过检验的系数序列，不能拒绝系数序列为白噪声序列的假设，则直接放弃该系数序列。然后对剩下的系数序列进行支持向量回归，首先将系数序列映射到 [0,1] 区间上，然后将已知的系数序列作为训练集，使用参数寻优算法，选择最合适的参数得到最优预测模型，对于给定的输入属性集，获得预测的系数值。

4. 最后通过 ARIMA 得到了趋势系数预测序列和通过 SVR 得到了细节系数预测序列，然后将各部分预测序列通过最优小波包重构得到最终的股价预测序列。

七、股票价格预测实证分析

实证分析数据来自于深证 A 股桑德环境(000826) 自 2001 年 1 月 1 日到 2010 年 1 月 1 日的日收盘价数据，共 2015 个样本观测值。桑德环境公司是一家在深圳证券交易所公开上市的国内环保龙头企业，其市值保持在较大的规模，同时成交量大，交易活跃，并且其各项盈利指标每年均保持稳定的快速增长。其信息披露公开透明，同时股价对公开披露信息反映迅速，其股价变动具有很强的规律性，因此很适合用来作为本文提出的股价预测模型实证分析的样本。对样本前 2000 个收盘价作为已知数据建立模型，并根据此模型预测未来 15 日的日收盘价。股票价格序列如图 4 所示。

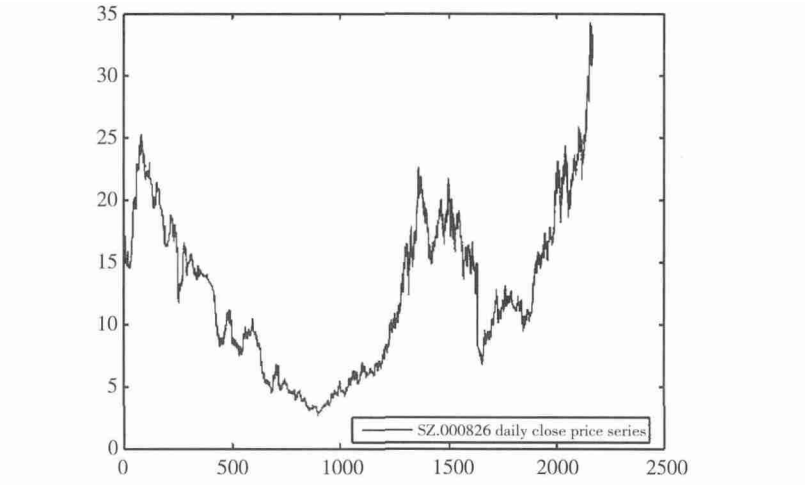

图 4　桑德环境 2001 至 2010 日收盘价

本文使用 MATLAB 的小波工具箱和计量工具箱进行小波变换与 ARIMA 回归，使用 LIBSVM 工具箱进行 SVR 回归预测。最后的预测结果使用 MAPE 进行衡量，MAPE 可表示为:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - P_i}{A_i}\right|$$

其中 A_i 为第 i 天的实际值，P_i 为第 i 天的预测值，总预测集大小为 n。

首先对股价序列求对数处理，然后使用 db4 小波对原始股价序列进行 4 层小波包分解，代价函数使用信息熵函数，得到了趋势系数 c_{15} 和细节系数集 c_2，c_{10}，c_{16}，c_{17}，c_{18}，c_{19}，c_{20}。

(一) 趋势系数的 ARIMA 回归

首先对趋势系数 c_{15} 进行 ARIMA 回归，c_{15} 的自相关图和偏自相关图如图 5 所示:

可见该序列具有很强的趋势性，因此尝试一阶差分，一阶差分后自相关图和偏自相关图如图 6:

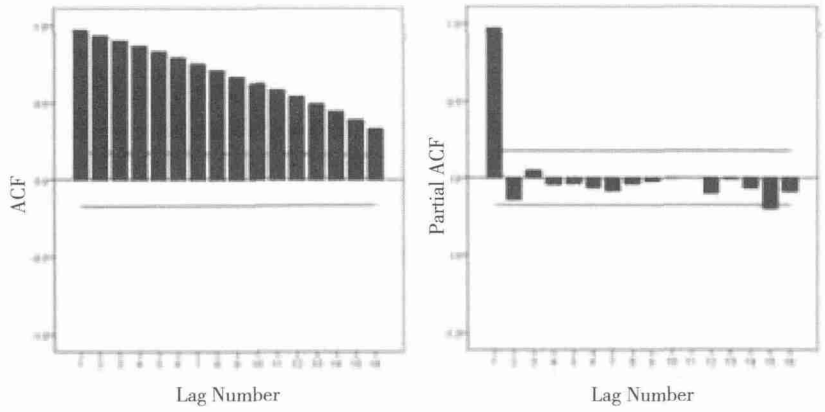

图 5 c_{15} 的自相关图和偏自相关图

左图纵坐标文本为 ACF，横坐标文本为 Lag Number；右图纵坐标文本为 Partial ACF，横坐标文本为 Lag Number

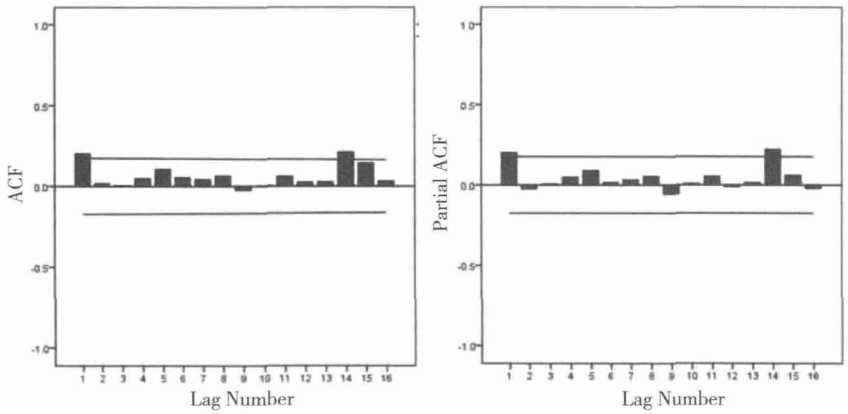

图 6 c_{15} 一阶差分后的自相关图和偏自相关图

左图纵坐标文本为 ACF，横坐标文本为 Lag Number；右图纵坐标文本为 Partial ACF，横坐标文本为 Lag Number

可见从一阶差分自相关图和偏自相关图并不能得到明确的定阶结果，实际上只有到三阶差分才可以定阶，三阶差分后自相关图和偏自相关图如图 7 所示：

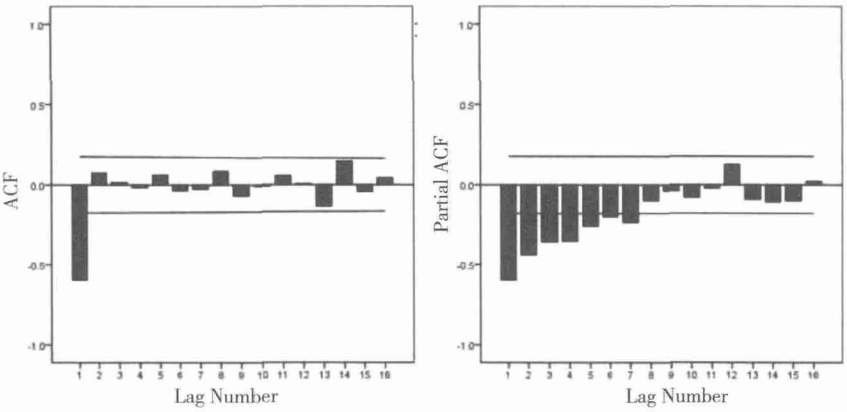

图 7 c_{15} 三阶差分后的自相关图和偏自相关图

左图纵坐标文本为 ACF，横坐标文本为 Lag Number；右图纵坐标文本为 Partial ACF，横坐标文本为 Lag Number

从图 7 中可以确定应当采用 ARIMA(4,3,1) 或者 ARIMA(5,3,1)，通过标准化 BIC 准则应当选择 ARIMA(4,3,1)。使用 ARIMA(4,3,1) 对趋势系数 c_{15} 进行拟合预测，拟合结果如表 1 所示:

表 1　　拟合残差 LBQ 检验

Model Fit	Ljung – Box Q(18)		
Stationary R – squared	Statistics	DF	Sig.
0.783	13.596	13	0.403

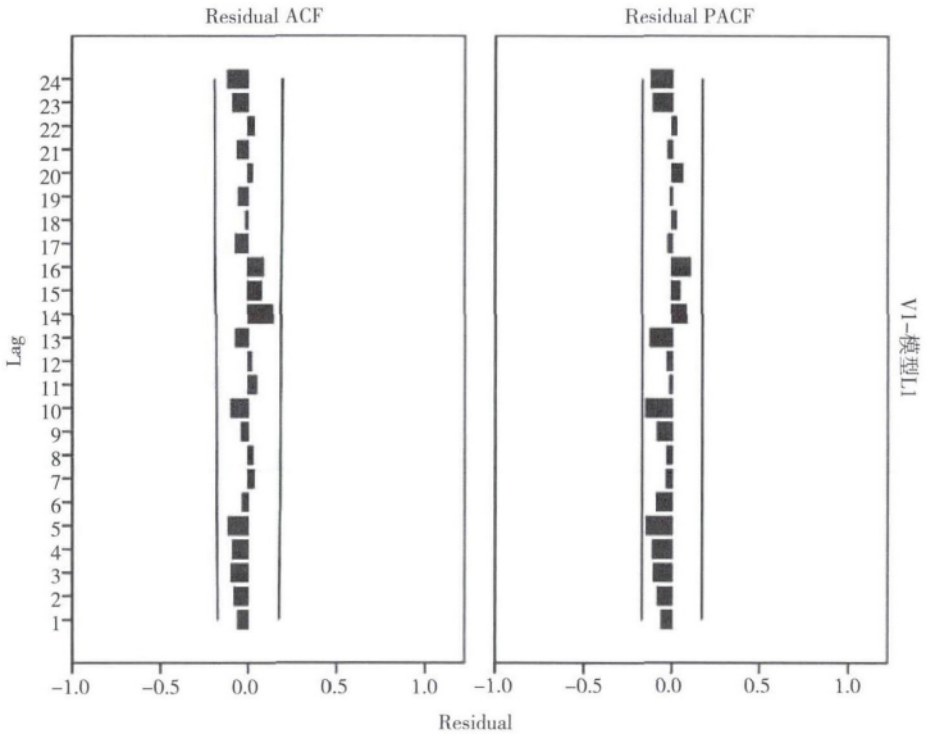

图 8　残差自相关图和偏自相关图

从表 1 和图 8 可知，拟合效果较好，残差通过了 LBQ 随机性检验，可以认为是白噪声序列，并且从残差自相关和偏自相关图也可以观察到趋势信息基本被提取完毕。

表 2　　系数估计值和显著性

		Estimate	SE	t	Sig.
AR	Lag 1	–0.616	0.094	–6.545	0.000
	Lag 2	–0.453	0.108	–4.197	0.000
	Lag 3	–0.359	0.109	–3.308	0.001
	Lag 4	–0.158	0.095	–1.663	0.099
Difference		3			
MA	Lag 1	0.997	0.500	1.994	0.048

从表 2 可发现前三阶滞后和一阶移动平均估计系数在 5% 的置信度下都较为显著，第四阶滞后估计系数在 10% 置信度下显著，拟合情况较好，拟合结果如图 9 所示，其中虚线为拟合值。

c_{15} 的 ARIMA 回归预测值为 4.5115，实际值为 4.3974。

(二) 细节系数的 SVR 回归

接下来对细节系数集进行预测，首先进行 LBQ 检验，对 c_{17}，c_{19}，c_{20} 进行检验时，不能拒绝是白噪声序

列的原假设，因此去除掉 c_{17}，c_{19}，c_{20}。对剩下的 c_2，c_{10}，c_{16}，c_{18} 进行三阶滞后 SVR 一步回归，将滞后三阶的观察值作为属性集，将已观察到的系数作为训练集。需要注意的是，在进行 SVR 一步预测时，往往需要得到最近的小波包系数序列，但这种情况下不能直接将待预测的序列部分也同时进行小波包分解，这样会使得已知的小波包系数带有未来几天对当前来说未知的信息，如果使用这样的小波包系数序列进行预测，会使得预测结果和实际股价极为一致，但实际上，由于受到经济环境、社会政策、投资者心理和公司运营的不确定性，股票市场具有很强的不可预知性，任何量化方法都达不到非常高的准确率，因此在任何一个预测时点上，只能根据已知的信息逐步进行小波包分解，才能避免当前的小波包系数带有未来信息，从而得到真实的预测值。第一步对数据归一化预处理，将系数序列映射到 [0，1] 区间。在进行 SVR 回归预测前需要选择核函数的形式，从实证预测精度效果来说，多项式核函数比其他核函数更加适合进行本文的回归。多项式核函数的形式为：

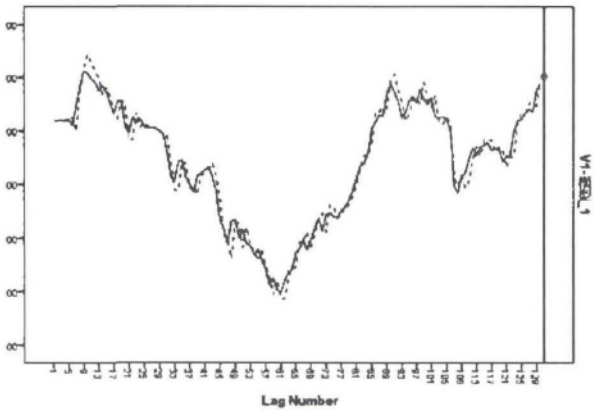

图 9 c_{15} 的 ARIMA 拟合结果图

$$K(x,y) = (x \cdot y + g)^d$$

对于 SVR 模型来说，决定模型精确度的重要一环是确定惩罚系数 C 和核函数 Gamma 系数，在本文中使用 PSO 粒子群优化算法进行参数寻优，PSO 粒子群优化算法利用个体的信息共享使整体问题求解空间从无序到有序的收敛过程，从而获得最优解。PSO 粒子群优化算法与 GA 遗传算法相比较而言，求解过程更为简单，但同时还能保持较好的精度。另外在参数寻优时同时使用 5 折 Cross—Validation 算法，进一步提高参数寻优效果。

在进行 SVR 预测时发现，直接使用系数序列进行训练和预测会产生“放弃预测”的现象，以细节系数 c_{18} 为例，如图 10 所示。

产生这一现象的主要原因是序列波动过于剧烈，变动趋势特征不明显，使得对于 SVR 来说做出预测的效果还不如放弃预测效果好。考虑到这种情况，对所有细节系数进行 5 阶滞后指数平滑法来提取特征。图 11 是处理后的拟合预测效果。

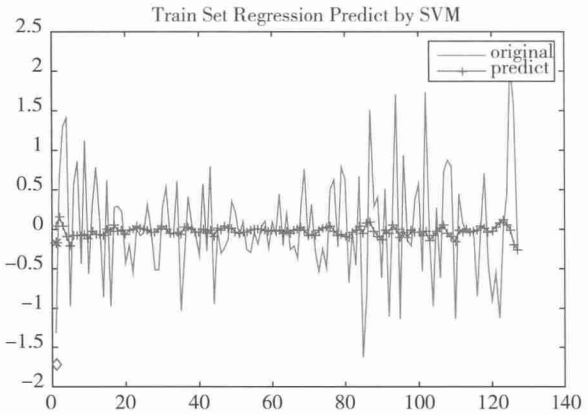

图 10 c_{18} 的 SVR 拟合预测效果图

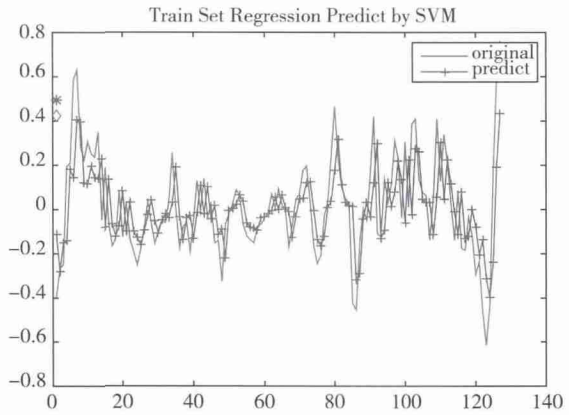

图 11 c_{18} 5 阶平滑后的 SVR 拟合预测效果图

从图 11 可以发现，在 5 阶滞后指数平滑后，SVR 能够有效地拟合和做出预测，对其他细节系数采用相同的处理方法，最终可以得到 c_2，c_{10}，c_{16}，c_{18} 的 SVR 预测序列，如图 12、13、14、15 所示。

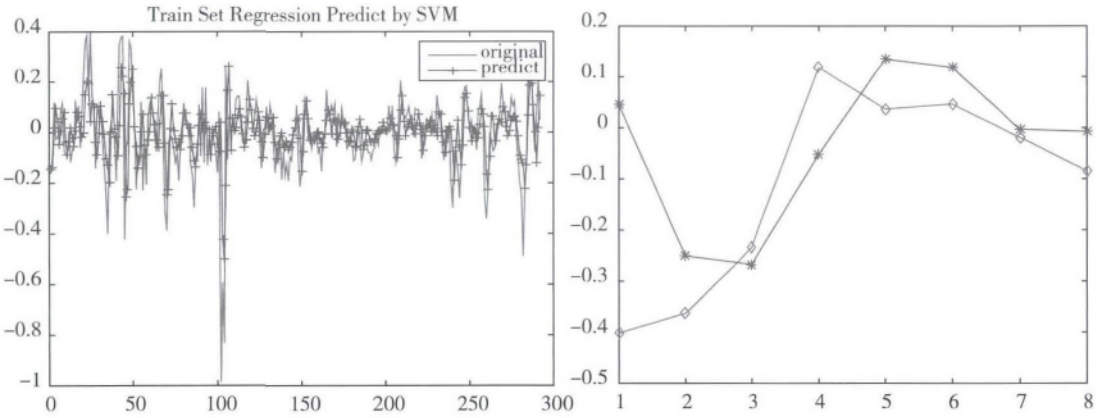

图 12　c_2 5 阶平滑后的 SVR 拟合预测效果图

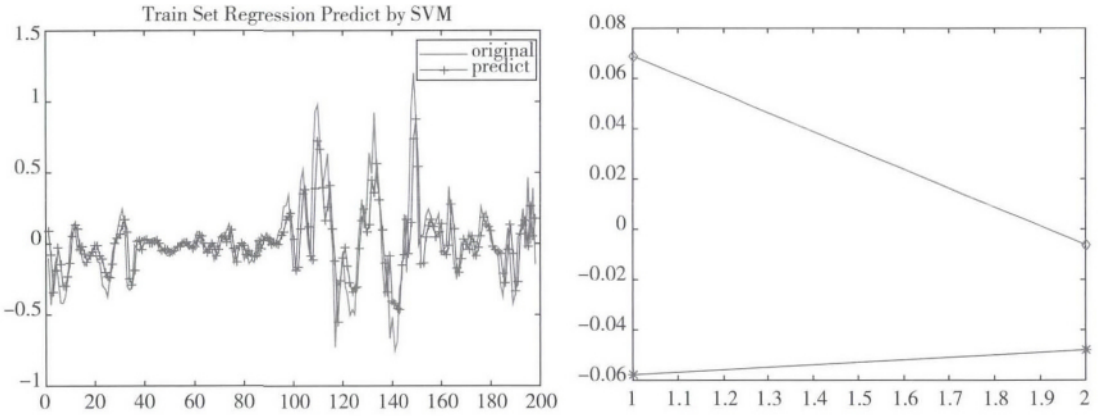

图 13　c_{10} 5 阶平滑后的 SVR 拟合预测效果图

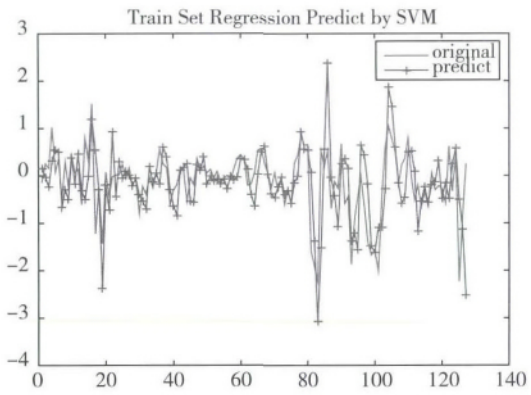

图 14　c_{16} 5 阶平滑后的 SVR 拟合效果图

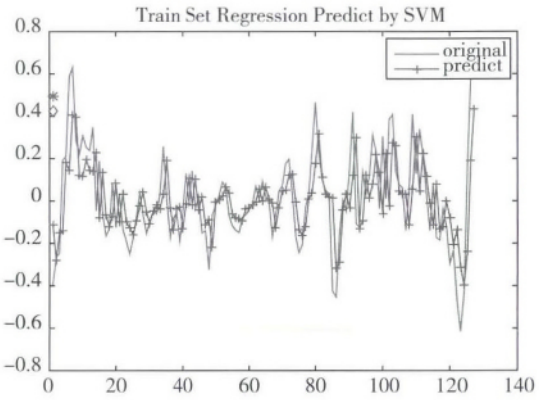

图 15　c_{18} 5 阶平滑后的 SVR 拟合预测效果图

c_{16} 5 阶平滑后的 SVR 预测值为 1.5316，实际值为 1.4851。

c_{18} 5 阶平滑后的 SVR 预测值为 0.1356，实际值为 −0.0162。

（三）小波包重构预测序列

通过 SVR 获得各个细节系数预测序列后，将各预测系数序列重构为股价预测序列。混合模型和 ARIMA 模型的未来 15 天的日收盘价预测结果对比图 16 所示，其中星号线段为实际日收盘价，三角号线段为本文提出的混合模型的预测日收盘价，十字号线段为 ARIMA 模型的预测日收盘价：

从表 3 最终的预测结果来看，总体预测较好，并且总体 MAPE 达到了 5.61%，并没有产生一些方法容易产生的“平移现象”和“放弃预测现象”，而 ARIMA 模型的 MAPE 为 4.31%，但是从图 16 可以很明显地看到 ARIMA 做出的预测有明显的“平移现象”，虽然其 MAPE 误差较低，但是预测效果很差。而如果从划分时间段来看，本文提出的混合模型在前 7 天的预测效果较好，其 MAPE 达到了 2.5%，而且从图形上来看，其预测值也基本与实际值吻合，ARIMA 在前 7 天则较差，其 MAPE 达到了 5.42%，且预测效果较差，7 天内只预测对了两天。观察后 8 天，混合模型预测的效果则稍差一些，其 MAPE 为 9.4%，但其走势方向还是一致的，这样的结果最主要的原因可能是对股价序列进行小波包分解后，趋势系数序列反映了股价变动的大致方向，但是该序列的预测值在四层小波包分解时大致覆盖了 15 天的价格序列，因此在比较临近的几天内是比较准确的，但是随着预测天数越来越靠后，这样的信息并不能有效反映相隔较远日期的趋势变化从而产生局部失真，而高频部分由于尺度较小，因此能够迅速地反映出近期的价格波动，因此在后半段的波动变化较为一致，而 ARIMA 仍然有明显的“平移现象”，还是没有做出有效预测。也就是说本文所提出的方法在预测未来 7 天以内的股价是较为可靠的，而在未来 8 到 15 天则可以较准确判断股价变动方向。

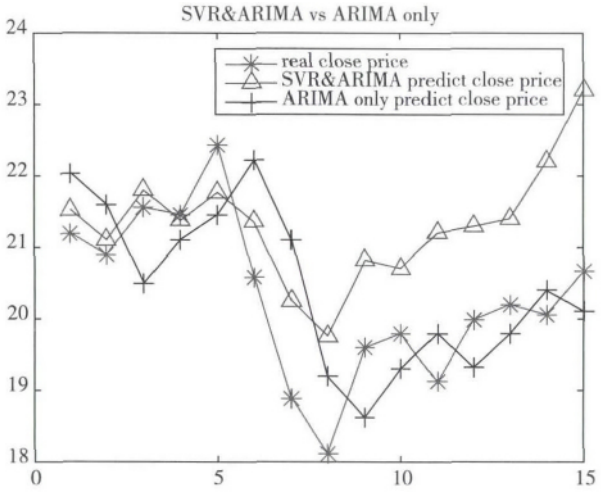

图 16 未来 15 天的日收盘价预测结果

表 3 模型最终预测结果

日期	实际值	混合模型预测值	混合模型 MAPE	ARIMA 模型预测值	ARIMA 模型 MAPE
1	21.200	21.526	15 天 MAPE	22.0400	15 天 MAPE
2	20.900	21.107		21.6000	
3	21.560	21.800	5.61%	20.5000	4.31%
4	21.460	21.377		21.1100	
5	22.430	21.771	前 7 天 MAPE	21.4600	前 7 天 MAPE
6	20.580	21.359		22.2300	
7	18.890	20.248	2.56%	21.1100	5.42%
8	18.120	19.762		19.1900	
9	19.600	20.820		18.6200	
10	19.800	20.706	后 8 天 MAPE	19.3000	后 8 天 MAPE
11	19.120	21.203		19.8000	
12	19.990	21.295		19.3200	
13	20.200	21.401	9.45%	19.7900	3.34%
14	20.060	22.200		20.4000	
15	20.670	23.211		20.1100	

八、结论

本文提出了一种基于最优小波包变换、ARIMA 和 SVR 的股票价格序列预测方法，这种方法首先对股票价格序列进行小波包分解，并根据信息熵代价函数提取了最优小波包基，然后对提取的趋势系数使用 ARIMA 模型进行回归预测，对提取的细节系数集使用 SVR 模型进行回归预测，然后将各系数预测序列通

过小波包重构合成为最终的股票价格预测序列。通过实盘股票桑德环境的日收盘价对方法的有效性进行了检验，实验结果表明该方法能够有效地进行中短期预测，并且避免了“平移现象”和“放弃预测现象”。

参考文献：

[1] 祁筠超. 基于 ARIMA 模型对恒生指数的实证分析 [J]. 经济师，2014(8)：108－110.

[2] 董博伦，徐东钰. 基于 ARIMA 模型的农产品类股价预测与分析 [J]. 现代商业，2015(3)：186－188.

[3] 张超. 基于误差校正的 ARMA－GARCH 股票价格预测 [J]. 南京航空航天大学学报(社会科学版)，2014(3)：43－48.

[4] 陈园园，刘俊，傅强. 基于 EMD 的神经网络股价预测方法 [J]. 新疆大学学报(哲学人文社会科学版)，2014(4)：6－11.

[5] 张浩，张代远. 基于三次样条权函数神经网络的股价预测 [J]. 计算机技术与发展，2014(6)：27－31.

[6] Weimin Ma，Yingying Wang，Ningfang Dong. Study on Stock Price Prediction Based on BP Neural Network [A]. Proceedings of 2010 IEEE International Conference on Emergency Management and Management Sciences(ICEMMS2010) [C]. 2010.

[7] 俞国红，杨德志，丛佩丽. ARIMA 和 RBF 神经网络相融合的股票价格预测研究 [J]. 计算机工程与应用，2013(18)：245－248.

[8] 程昌品，陈强，姜永生. 基于 ARIMA－SVM 组合模型的股票价格预测 [J]. 计算机仿真，2012(6)：343－346.

[9] 杜建卫，王超峰. 小波分析方法在金融股票数据预测中的应用 [J]. 数学的实践与认识，2008(7)：68－75.

[10] 张坤，郁湧，李彤. 基于小波和神经网络相结合的股票价格模型 [J]. 计算机工程与设计，2009(23)：5496－5498.

[11] 隋学深，齐中英. 基于多尺度特征和支持向量机的股市趋势预测 [J]. 哈尔滨工业大学学报(社会科学版)，2008(4)：77－81.

[12] 常松，何建敏. 基于小波包和神经网络的股票价格预测模型 [J]. 中国管理科学，2001(5)：8－15.

[13] 孙延奎. 小波分析及其应用 [M]. 北京：机械工业出版社，2005：245－257.

[14] Daubenchies I. Ten Lectures on Wavelet [M]. Pennsylvania：Capital City Press，1992：105－114.

[15] 薛薇. SPSS 统计分析方法及应用 [M]. 北京：电子工业出版社，2009：41－159.

[16] 邓乃扬，田英杰. 支持向量机：理论、算法与拓展 [M]. 北京：科学出版社，2009：63－111.

Research on Stock Price Prediction Based on Optimal Wavelet Packet Transformation and ARIMA－SVR Mixed Model

GAO Tian

(School of Insurance，Central University of Finance and Economics，Beijing 100081，China)

Abstract: The changes in stock price series are always non－stationary and heteroscedastic，that fact makes single prediction method difficult to predict accurately. In this paper，using the best wavelet packet transform，the stock price series is decomposed into series of wavelet packet coefficients which reveal characteristics more obviously. The prediction of the trend coefficients using ARIMA Model，and the prediction of the detail coefficients using SVR Model，the predicted results are reconstructed to obtain the stock price forecasting sequence. The empirical result shows that，this mixed prediction method is of clear structure，high computation efficiency，and be able to predict the stock price changes with high accuracy.

Key words: optimal wavelet packet transformation；ARIMA；SVR；stock price；prediction

责任编辑：萧敏娜