文章编号：1006－9348(2010)05－0264－03

ARIMA与 SVM组合模型的石油价格预测

吴 虹，尹 华

(赣南师范学院数学与计算机科学学院，江西 赣州 341000)

摘要：针对复杂时间序列预测困难的问题，在综合分析其线性和非线性复合特征的基础上，提出了一种基于 ARIMA和 SVM相结合的时间序列预测模型。首先采用 ARIMA模型对时间序列进行线性建模，然后采用 SVM对时间序列的非线性部分进行建模，最后得到两种模型的综合预测结果。将组合模型应用于石油价格预测中，仿真结果表明组合模型相对于单模型的预测具有更高的精度，发挥了 2种模型各自的优势，在复杂时间序列预测中具有广泛的应用前景。

关键词：支持向量机；差分自回归移动平均；组合预测；石油价格

中图分类号：TP309 文献标识码：A

Oil Price Forecasting Based on ARIMA and SVM Hybrid Model

WU Hong, YIN Hua

(Gannan Normal University, Faculty of Mathematics and Computer, Ganzhou Jiangxi 341000, China)

ABSTRACT: In order to solve the problem of complex time series forecasting including the linear and nonlinear features, a new hybrid forecasting model based on ARIMA and SVM is proposed in this paper. ARIMA model was used to predict the linear component of complex time series and SVM model was applied to the nonlinear residual component and the hybrid forecasting results were obtained. The prediction performances of the methods are tested on simulation experiment for oil price. The results show that the hybrid model, which takes advantage of the unique strength of the two models in linear and nonlinear modeling, has better accuracy than the single model. The hybrid model is an effective method for complex time series.

KEYWORDS: Support vector machine (SVM); ARIMA; Hybrid forecastion; Oil price

1 引言

石油价格数据是一种高度不稳定、复杂且难以预测的时间序列数据，因为这些数据往往既隐含大量的动态特征，又受自变量的影响，同时具有高度的非线性。目前，石油期货价格预测中较多使用数量化模型，包括时间序列方法[1]、回归分析法[2]和人工神经网络方法[3]。最具代表性的时间分析方法为差分自回归移动平均(autoregressive integrating moving average, ARIMA)，ARIMA是基于线性数据的模型，其无法捕捉非线性数据的信息。而现实中时间序列数据间往往更多地表现为非线性的且含有复杂的噪声，这样基于线性模型定阶获得的模型阶数和保留变量的 ARIMA模型，往往并非最优，从而导致预测精度不高。神经网络等模型在非线性时序列序列预测中表现出其优越性[4]，然而，在实际应用中，神经网络学习算法表现出其不足：如隐含层数的选择、过拟合问题、局部极小值以及泛化性能不强。基于结构风险最小化的支持向量机(Support Vector Machines, SVM)，是一种新的机器学习方法，其在非线性时间序列领域里取得了不错的预测结果，较好地解决了小样本、非线性、过拟合、维数灾和局极小等问题，且泛化推广能力优异[5]。基于著名的 M—竞争理论[6]，为了有效地利用各种模型的优点，一些学者利用组合预测方法来进行时间序列预测研究[7]，实证结果表明，相对于单独的各种模型，组合模型比单个预测模型考虑问题更系统、更全面，因而能够有效地减少单个预测模型过程中一些环境因素的影，从而提高预测的精度。

目前，通过 ARIMA和 SVM组合预测方法进行石油价格的预测还鲜有文献，本文提出了一种基于 ARIMA和支持向量组合模型的石油价格预测新方法— ARIMA— SVM。ARIMA模型描述历史数据的线性关系，SVM模拟数据的非线性规律。对石油价格进行仿真实验，验证 ARIMA— SVM模型的有效性和可行性。

2 ARIMA— SVM模型

2.1 ARIMA— SVM原理

ARIMA是一种精确度较高的线性时间序列预测方法，时间序列分析是处理动态数据的一种有效的参数化时域分

基金项目：国家自然科学基金(30570352)资助。

收稿日期：2009－12－30 修回日期：2010－01－14

析方法，是 20世纪 70年代美国学者鲍克斯·乔瑞（George Box）和英国统计学家詹肯·格威勒姆（Gwilym Jenkins）所建立的鲍克斯—詹姆（B—J）方法的进一步发展和改进[8]。支持向量机是 Vapnik等[9]根据统计学理论提出的一种新的通用学习方法，它是建立在统计学理论的 VC维理论和结构风险最小化原理基础上，能较好地解决小样本、非线性、高维数和局部极小点等实际问题，被视为替代人工神经网络的较好算法。

由于石油市场具有复杂的非线性动力系统特征，它既受确定性规律支配，同时又表现出某种随机现象，即石油价格具有时变性、随机性和模糊性的特征，所以单纯使用 SVM或 ARIMA模型进行预测都有可能导致误差过大[10]。因此，可以先使用 ARIMA模型预测石油价格历史数数据，使其线性规律信息包含在 ARIMA模型的预测结果中，这时非线性规律包含在了 ARIMA模型的预测误差中。然后用 SVM预测 ARIMA模型的误差，使非线性规律包含在 SVM的预测结果中。最后用 ARIMA的预测结果与 SVM的预测相加得到组合预测模型的预测值，其原理如图 1所示。

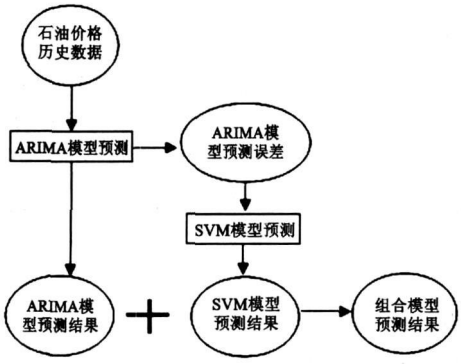

图 1　ARIMA和 SVM的组合预测模型流程图

2.2　ARIMA—SVM算法具体步骤

把一组时间序列的数据 y_t看成是由非线线自相关结构 L_t和非线性结构 N_t两部分，即

$$y_t = L_t + N_t \tag{1}$$

具体建模步骤如下：

步骤 1，用 ARIMA模型对 y_t进行预测。设预测结果为 $\hat{L}_t$，原序列与 ARIMA模型预测结果的残差为 e_t，即

$$e_t = y_t - \hat{L}_t \tag{2}$$

序列 $\{e_t\}$是隐含了原序列中的非线性关系

$$e_t = f(e_{t-1}, e_{t-2}, \cdots, e_{t-n}) + \varepsilon \tag{3}$$

式中 ε为随机误度。

步骤 2　根据步骤 1得到的残差序列，利用步骤 1的 ARIMA模型的阶数确定输入残差阶数，重构残差时间序列样本，利用 SVM对残差进行预测，并设预测结果为 $\hat{e}_t$。

步骤 3　利用 2种模型的测结果组合成为最终的预测结果 $\{\hat{y}_t\}$。结果为

$$\hat{y}_t = \hat{L}_t + \hat{e}_t \tag{4}$$

2.3　参比模型及评价指标

为了考察 ARIMA—SVM模型与 SVM模型、ARIMA模型的优劣，所有模型采用一步预测法。ARIMA由 DPS6.55给出，SVM由自编 MATLAB7.0通过调用 LIBSVM2.8实现，数据处理技术与 ARIMA—SVM相同。为了评价模型预测性能的优劣，使用均方根误差（Root Mean Square Error，RMSE）和 MAPE作为模型的评价指标。RMSE和 MAPE分别定义如下：

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{i}(y_i - \hat{y}_i)^2} \tag{5}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{i}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{6}$$

其中，y_i为分别为真值，和 $\hat{y}_i$为预测值，n为预测样本数。RMSE仅适用于同一数据集不同模型间的比较，MAPE可用于不同数据集间的比较。但对同一数据集，如 A模型与 B模型相比虽 MAPE较大而 RMSE较小，则 A模型预测更为稳健，因此，RMSE为主要评价指标。

2.4　支持向量机软件

目前支持向量机软件的很多，最为流行的为台湾国立大学林智仁教授的 LIBSVM，LIBSVM算法是一种将序列最小优化算法（Sequential Minimal Optimization，SMO）和 Svmlight算法相结合的优化方法，对工作集的选择策略进行了改进。其含四个常用程序：svm—scale用于对原始数据规格化，svm—train用于训练，svm—predic用于预测，gridregression.py用于自动搜索核函数最优参数。

3　ARIMA—SVM模型仿真实验

3.1　数据来源

采用西德克萨斯原油指数（WTI）1986年 1月至 2009年 9月的月度统计数据（单位：美元/桶）作为预测世界石油价格的实证研究。数据源是：U.S. Energy Information Administration　http://tonto.eia.doe.gov/dnav/pet/hist/rwtcd.htm，石油价格时间序列走势如图 2所示。其中 1986—2007年的数据作为训练集，2007—2009年的数据作为验证集，2009.01—2009.09的数据作为测试集，来验证组合模型的有效性。

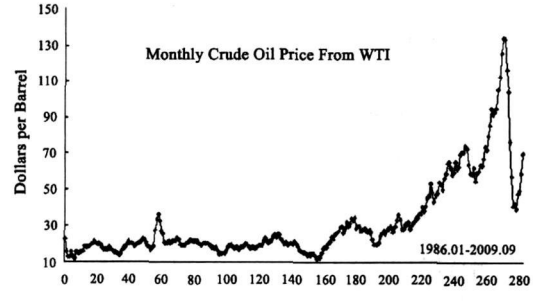

图 2　石油价格走势图（WTI）

3.2 模型的实现

3.2.1 石油价格预测的 ARIMA模型建立过程

1)石油价格历史数据的平稳化

从图 2可以看出入石油价格的变动呈现明显的非线性特征,不仅局部变动剧烈,在大尺度的时间范围内也是跌宕起伏,这表示该时间序列存在方差不齐,因此,需要对其进行平稳化处理。对数据分别了一次差分和自然对数一次差分,发现自然对数一次差分后后已经基本平稳化,所以设定 ARIMA模型参数 d= 1。

2)ARIMA模型 p和 q的确定及预测

首先,借助 DPS6.55构建 ARIMA模型,采用从低阶到高阶逐步试探法来识别模型的类型和阶数,经过比较分析,选择 ARIMA(1,4,0)模型,预测结果见图 3发现其拟合效果较好,可以进行石油价格预测。预测样本的 MSE为 23.63,可见 ARIMA模型得到了较为好的预测效果。

3.2.2 石油价格非线性 SVM建模过程

1)模型的定阶

模型的阶数也称为嵌入维数,其选取方法在实践中有两种,一是靠经验选择,二是先设定模型其他参数,然后对滞后阶数按照一定的标准进行优化。第一种方法过分依赖研究者的水平和经验,不能客观选择最佳滞后阶数,而第二种方法则忽略了一个重要问题:滞后阶数与其他参数对模型好坏的影响是相互的,如果先人为设定其他参数选择滞后阶数,而后使用该滞后阶数值优化参数,很有可能只是在该滞后阶数下的参数最优,而不是全局最优。本文将模型阶数与 SVM模型参数一起寻优,具体如下:

假定一多输入单输出回归模型有 N个样本、一个因变量、m-1个自变量,由低阶到高阶递增地以 SVM进行留一法测试,并依 RMSE最小标准决定拓展阶数与否。对待比较的相邻两模型 SVM(n)和 SVM(n+1),记 $RMSE_{SVM(n)}$为 SVM(n)的均方误差,$RMSE_{SVM(n+1)}$为 SVM(n+1)的均方误差。若 $RMSE_{SVM(n)} > RMSE_{SVM(n+1)}$,继续拓阶;若 $RMSE_{SVM(n)} \leqslant RMSE_{SVM(n+1)}$,拓阶终止,取 SVM(n)为定阶后模型。最后模型的阶数为 4。

(2)序列的重构及预测

从模型定阶可知当月石油价格的残差受到前 4个月石油价格残差重要的影响,这就意味着将前 4个月石油价格的残差作为 SVM的输入来预测当月石油价格的残差。在 MATLAB7.0平台下自编程调用 LIBSVM工具箱来实现 SVM建模,核函数为高斯核函数,采用 10折交叉验证,经 gridregression.py自动搜索确定模型最优参数,最优参数为:径向基宽度 γ= 0.5,ε= 0.25,C= 512,使用最优参数下训练得到的模型对 2009.01-2009.09石油价格残差进行预测。

3.2.3 组合模型的验证结果

最后,根据 ARIMA模型得到的线性预测结果和 SVM模型的非线性预测结果,进行简单的相加得到 ARIMA-SVM模型的预测结果。各模型的预测结果如图 3所示,RMSE和 MAPE见表 1。

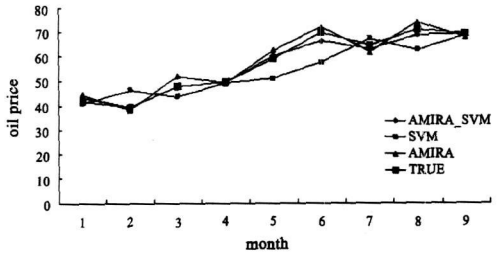

图 3各种模型的预测结果与实际原油价格(2009.01-09)结果对比

表 1 各种预测方法的误差比较

预测法	RMSE	MAPE
ARIMA	23.63	0.03627
SVM	6.29	0.0204
ARIMA-SVM	1.60	0.00455

3.3 结果分析

从图 3,表 1中可以看出,非线性的 SVM模型比线性的 ARIMA模型的预测效果好;而 ARIMA-SVM模型与 ARIMA和 SVM相比,预测精度有很大的提高,具有明显的优势。因此,从实验结果验证了 ARIMA-SVM预测模型的有效性和可行性。

4 结论

近些年来,石油价格的分析和预测是一个非常活跃的研究领域,其具有线性关系和非线性特征,由于 ARIMA和 SVM模型分别对线性和非性性问题有其相对的优势,但它们对于复杂的、不稳定的时间序列都不是最优的模型。本文利用 ARIMA模型捕捉石油价格中的线性趋势,用 SVM预测石油价格的非线性规律,形成组合预测模型。通过对 WTI油价的实证研究表明了预测模型在长期预测上的有效性,能够总体把握油价的趋势,达到更准确地对世界油价进行预测的目的,验证了组合模型比单一模型的预测结果更合理、更可靠,该预测模型是一种有效的石油价格时间序列预测模型。

参考文献:

[1] W M Fong, K H See. A markov switching model of the conditional volatility of crude oil futures prices[J]. Energy Economics, 2002, 24: 71-95.

[2] Imad A Moosa, Param Silvapulle. The price volume relationship in the crude oil futures market: some results based on linear and non-linear causality testing[J]. International Review of Economics and Finance, 2000, 9: 11-30.

(下转第 326页)

16种可能的排列，统计这 16种排列出现次数，用 f(i)(0≤ i ≤ 15)表示。计算 X=(16/5000) * {$\sum_{i=0}^{15}$[f(i)] 2} - 5000，如果 2 16< X< 46 17，表明通过检验。

游程被定义为连续的 0或者 1的最大序列，游程是序列的子串，其前后元素都与其本身元素不同，游程检验主要检验序列中游程总数是否符合随机性要求，如果游程总数满足一定的范围，则认为通过检验，所有大于 6的游程都被认为长度为 6。

对于长度大于或等于 26的游程就认为是长游程，在 20000比特里，如果没有出现长游程，就认为通过检验。

具体结果与通过率如表 1所示，由表 1可以看出所有的测试都通过。

表 1 FIPS140-2测试结果

测试	下限值	上限值	测试结果	通过率
频数检验	9725	10275	10008	100%
扑克检验	2 16	46 17	16 75	100%
1游程检验	2315	2685	2498	100%
2游程检验	1114	1386	1281	100%
3游程检验	527	723	636	100%
4游程检验	240	384	308	100%
5游程检验	103	209	157	100%
6游程检验	103	209	139	100%
长游程检验	—	—	0	100%

5 结论

本文在深亚微米工艺下设计了一种物理噪声源，输出速率为 1.5Mbps。采用瞬态噪声仿真方法，使得在仿真阶段就可以对电路生成的比特流进行随机性分析。由此得到的数字序列通过了 FIPS140-2国际标准测试。随着噪声成为深亚微米电子学中需要考虑的一个重要因素，即使瞬态噪声仿真方法的计算复杂度较高，但它仍将在仿真电路设计中有广泛的应用。

参考文献：

[1] M Gupta. Applications of Electrical Noise[C]. Proc IEEE, July 1975 (63): 996-1010.

[2] W T Hofman, J A Connelly, A B Dowlatabadi. An integrated analog/digital random noise source[J]. IEEE Transactions on Circuits and Systems, I, Jun 1997, 44(6): 521-528.

[3] G Denk, R Winkler. Modelling and simulation of transient noise in circuit simulation[J]. Mathematical and Computer Modelling of Dynamical Systems, August 2007, 13(4).

[4] Craig S Petrie and J A Alvin Connelly. A noise-based random bit generator IC for applications in cryptography[J]. School of Electrical and Computer Engineering Georgia Institute of Technology Atlanta, GA 30332-0250 USA, 1998.

[5] M Bucci. A High Speed Oscillator Based True Random Number Generator for Cryptographic Applications on a Smart Card IC[J]. IEEE Trans Computers, April 2003, 52(4): 403-409.

[6] J J Sung, G S Kang and S Kim. A transient noise model for frequency-dependent noise sources[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Aug 2003, 22(8).

[7] R J Baker, H W Li and D E Boyce. CMOS-Circuit Design, Layout and Simulation[M]. IEEE Press, 1998.

[作者简介]

江 军(1983-)，女(汉族)，安徽人，硕士研究生，研究方向为集成电路设计；

段成华(1962-)，男(汉族)，重庆人，教授，硕士研究生导师，研究方向为大规模集成电路设计，特种密码芯片设计。

(上接第 266页)

[3] Zhang Jia-shu, Li Heng-chao, Xiao Xian-ci. ADCT domain quadratic predictor for real-time prediction of continuous chaotic signal[J]. Acta Physical Sinica, 2004, 53(3): 710-716.

[4] L P Maguire, B Roche, T M Mcginnity. Predicting a chaotic time series using a fuzzy neural network[J]. Information Sciences, 1998, 112: 125-136.

[5] L P Maguire, B Roche, T M Mcginnity. Predicting a chaotic time series using a fuzzy neural network[J]. Information Sciences, 1998, 112: 125-136.

[6] Iulian B Ciocoiu. Time series analysis using RBF network swith FIR/IIR synapses[J]. Neurocomputing, 1998, 20: 57-66.

[7] Francis E H Tay, Cao Li-juan. Application of Support Vector Machines in Financial Time Series Forecasting[J]. The International Journal of Management Science, 2001, 29: 309-317.

[8] G P Zhang. Time series forecasting using a hybrid ARIMA and neural network model[J]. Neurocomputering, 2003, 50: 159-175.

[9] V Vapnik. The Nature of Statistical Learning Theory[M]. New York: Springer-Verlag, 1999.

[10] 冯春山，吴家春，蒋馥. 石油价格的组合预测研究[J]. 石油大学学报(社会科学版)，2004, 20(1): 12-14.

[作者简介]

吴 虹(1983-)，女(汉族)，江西宁都人，硕士，助教，主要研究方向：计算机应用；

尹 华(1975 8-)，女(汉族)，江西大余人，硕士，讲师，主要研究方向：计算机应用。